Systems/Circuits

# A Redundant Cortical Code for Speech Envelope

Kristina B. Penikis[1] and Dan H. Sanes[1,2,3,4]

[1]Center for Neural Science, New York University, New York, New York 10003, [2]Department of Psychology, New York University, New York, New York 10003, [3]Department of Biology, New York University, New York, New York 10003, and [4]Neuroscience Institute, New York University Langone Medical Center, New York University, New York, New York 10016

Animal communication sounds exhibit complex temporal structure because of the amplitude fluctuations that comprise the sound envelope. In human speech, envelope modulations drive synchronized activity in auditory cortex (AC), which correlates strongly with comprehension (Giraud and Poeppel, 2012; Peelle and Davis, 2012; Haegens and Zion Golumbic, 2018). Studies of envelope coding in single neurons, performed in nonhuman animals, have focused on periodic amplitude modulation (AM) stimuli and use response metrics that are not easy to juxtapose with data from humans. In this study, we sought to bridge these fields. Specifically, we looked directly at the temporal relationship between stimulus envelope and spiking, and we assessed whether the apparent diversity across neurons' AM responses contributes to the population representation of speech-like sound envelopes. We gathered responses from single neurons to vocoded speech stimuli and compared them to sinusoidal AM responses in auditory cortex (AC) of alert, freely moving Mongolian gerbils of both sexes. While AC neurons displayed heterogeneous tuning to AM rate, their temporal dynamics were stereotyped. Preferred response phases accumulated near the onsets of sinusoidal AM periods for slower rates (<8 Hz), and an over-representation of amplitude edges was apparent in population responses to both sinusoidal AM and vocoded speech envelopes. Crucially, this encoding bias imparted a decoding benefit: a classifier could discriminate vocoded speech stimuli using summed population activity, while higher frequency modulations required a more sophisticated decoder that tracked spiking responses from individual cells. Together, our results imply that the envelope structure relevant to parsing an acoustic stream could be read-out from a distributed, redundant population code.

*Key words:* amplitude modulation; auditory; cortex; envelope; speech; temporal coding

**Significance Statement**

Animal communication sounds have rich temporal structure and are often produced in extended sequences, including the syllabic structure of human speech. Although the auditory cortex (AC) is known to play a crucial role in representing speech syllables, the contribution of individual neurons remains uncertain. Here, we characterized the representations of both simple, amplitude-modulated sounds and complex, speech-like stimuli within a broad population of cortical neurons, and we found an overrepresentation of amplitude edges. Thus, a phasic, redundant code in auditory cortex can provide a mechanistic explanation for segmenting acoustic streams like human speech.

## Introduction

Animal communication sounds, including human speech, are produced in extended sequences of "syllables," or packets of information. The temporal structures of natural sounds are perceptually and behaviorally informative (Singh and Theunissen, 2003; Peelle and Davis, 2012; Haegens and Zion Golumbic, 2018). Studies in humans have established that slower rates of amplitude modulation (AM), below ~8 Hz, are correlated with successful speech reception (Drullman et al., 1994a, b; Smith et al., 2002; Ghitza, 2012). Fluctuations in sound envelope drive rhythmic activity in auditory cortex (AC), and the fidelity of this signal correlates closely with speech comprehension (Giraud and Poeppel, 2012; Peelle and Davis, 2012; Haegens and Zion Golumbic, 2018). Rhythmic tracking (sometimes called entrainment) is thought to reflect the neural mechanism that parses the acoustic stream into segments, which are subsequently used during phonemic processing. Much of what is known about the neuronal mechanisms of envelope processing in AC, however, comes from responses to simplified, periodic AM stimuli, assessed through single-unit (SU) recordings in animal models. Here, we asked how

Received Aug. 6, 2021; revised Aug. 19, 2022; accepted Oct. 23, 2022.

Author contributions: K.B.P. and D.H.S. designed research; K.B.P. performed research; K.B.P. analyzed data; K.B.P. wrote the first draft of the paper; K.B.P. and D.H.S. edited the paper; K.B.P. and D.H.S. wrote the paper.

This work was supported by National Institutes of Health Grants F31DC014903 (to K.B.P.) and R01DC011284 (to D.H.S.). We thank Shihab Shamma and Yulia Oganian for providing code that contributed to the analyses in Figures 7 and 9.

K. B. Penikis' present address: Department of Psychology, West Virginia University, Morgantown, WV 26505.

The authors declare no competing financial interests.

Correspondence should be addressed to Kristina B. Penikis at kbp254@nyu.edu.

https://doi.org/10.1523/JNEUROSCI.1616-21.2022

the temporal coding properties of single AC neurons contribute to speech envelope processing.

Single-cell studies have established that AC neurons display heterogeneous responses to AM rate (Schreiner and Urbas, 1988; Eggermont, 1998; L. Liang et al., 2002; Joris et al., 2004; Malone et al., 2007, 2013; Zhou and Wang, 2010; Yin et al., 2011; Hoglen et al., 2018). At the low AM rates discussed above, many neurons display phasic firing. Qualitatively, mean phases have been observed to vary between cells (Joris et al., 2004). A recent study of single units in squirrel monkeys using sinusoidal AM stimuli reported that many cortical neurons showed similar phase preferences and predicted that population coding of speech envelopes would be robust to indiscriminate pooling of cortical responses (Downer et al., 2021). A primary goal of the current study was to quantify the temporal relationship of spiking responses to both simple AM and complex speech-like envelopes and to investigate these encoding patterns in the context of decoding models.

To do so, we recorded from single AC neurons in awake gerbils and compared sinusoidal AM responses to those of vocoded speech stimuli. Although AC neurons were sampled across the tonotopic axis and were heterogeneous in terms of AM rate tuning, their temporal dynamics were relatively stereotyped. Spiking was biased toward the onsets of AM periods, illustrated by a majority of cells showing mean phases falling in the first 90° of a sinusoidal period. The encoding bias in individual neurons translated to an over-representation of amplitude edges in the gross population signal. For all speech-like stimuli, a transient, coherent population response was observed at onset edges. If a global neural signal sufficiently represents the envelopes of complex sounds like speech, then the marginal benefit of tracking individual neurons would be relatively small. Indeed, stimuli evoking this onset-biased encoding pattern were successfully decoded from a single-trial population-averaged activity vector, whereas stimuli without syllable-like envelopes required a more sophisticated decoding strategy for classification. Taken together, the results suggest that the temporal structure of a continuous sound stream, like speech, could be sufficiently captured by a redundant, population-level signal.

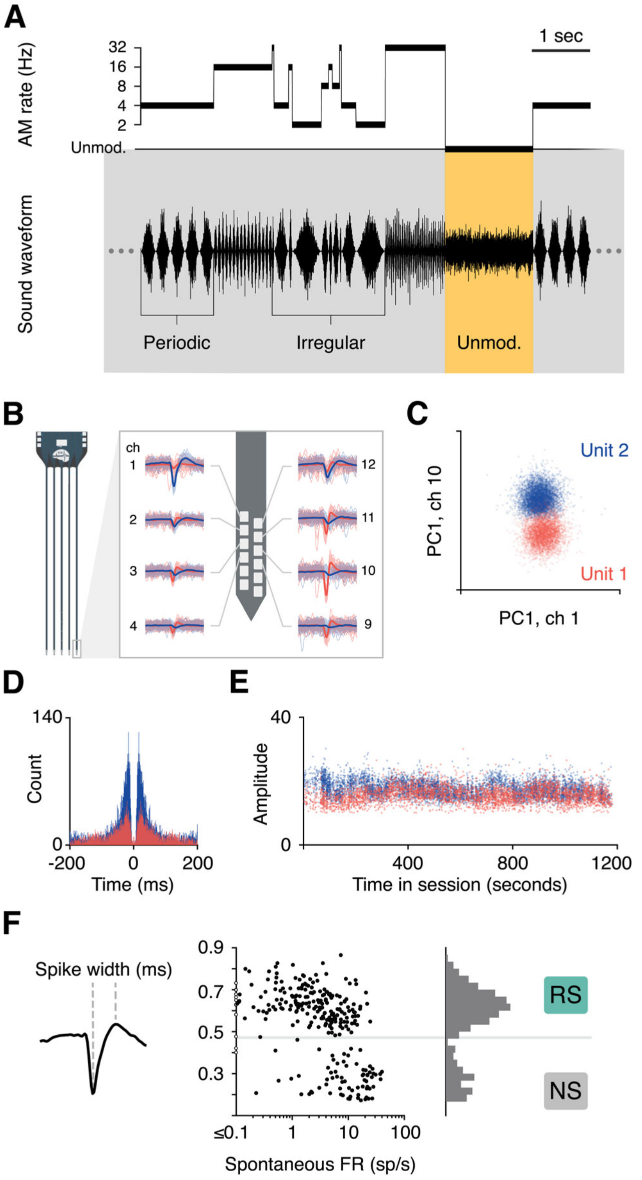

**Figure 1.** Experimental design and methodology. ***A***, The sinusoidal amplitude modulation (AM) stimulus consisted of a continuous broadband noise for which amplitude was either modulated or unmodulated. AM segments fluctuated periodically, at a single rate from 2 to 32 Hz, or irregularly, with individual periods from the same range presented in pseudorandom order. Gerbils performed a detection task in which they could safely drink water from a metal spout during all AM sounds. Unmodulated noise predicted a small electrical shock, which subjects learned to avoid by withdrawing from the spout momentarily. ***B***, Neural activity was recorded in auditory cortex during task engagement, and single units (SUs) were isolated offline. Waveforms of two units are shown on one shank of a 64-channel probe. ***C***, The two units were separated in principle component (PC) space. ***D***, The autocorrelation of spike times for each unit was confirmed to show a clean refractory period. ***E***, The distributions of waveform amplitudes across the session were inspected to ensure that a majority of spiking events were captured and that drift did not degrade unit quality. If all conditions were met, the unit was labeled as a SU and included in further analyses. ***F***, The width of spike waveforms for all SU resulted in a clean bimodal distribution, which was used to separate regular/broad spiking (RS) cells from narrow spiking (NS) cells.

## Materials and Methods

### Subjects

Electrophysiological data were obtained from adult Mongolian gerbils (*Meriones unguiculatus*; *N* = 5, 1 female). Animals were weaned at postnatal day 30 from commercial breeding pairs (Charles River) and housed on a 12/12 h dark/light cycle. Procedures related to the care and use of animals were approved by the Institutional Animal Care and Use Committee at New York University.

### Stimuli

For the sinusoidal condition, the acoustic stimulus consisted of a continuous stream of sinusoidally amplitude-modulated (AM) noise. AM depth was 100%, and the modulation rate was switched every 1–2 s, always at the trough/zero phase. Modulation rates were either periodic at an AM rate of 2, 4, 8, 16, or 32 Hz, or irregular: quasi-random sequences of individual periods drawn from this set of rates (Fig. 1*A*). The carrier was white noise filtered with a high pass cutoff of 100 Hz and a low pass cutoff of 1, 5, 10, 20, or 45 kHz, and mean level was either 45 or 60 dB SPL. These parameters were adjusted via a brief characterization protocol performed at the beginning of each session.

Vocoded speech stimuli were delivered to animals either after a behavioral session or on off-days (behavioral protocols discussed in section below). Excerpts from Seuss books were recorded in a sound-proof booth. Speech envelopes were extracted by low pass filtering (cutoff at 45 Hz) of the Hilbert transform. These envelopes were clipped into six stimuli, ranging from ∼1.5 to 6 s in duration, and then used to modulate white noise with carrier and level parameters set to match those presented in the immediately preceding behavioral session. Vocoded speech stimuli were presented pseudo-randomly in trial format with a 1-s interval from offset to onset. Several analyses used 500-ms tokens extracted from the full AM and vocoded speech stimuli. For more details, see below, Population poststimulus time histograms (PSTHs).

Gerbils were in a plastic testing cage within a sound-attenuating booth (Industrial Acoustics) and observed through a webcam. Sounds were presented from a calibrated tweeter (DX25TG05-04, Vifa) installed 1 meter directly above the testing cage. Calibration was performed using a one-quarter inch free-field condenser microphone (Bruel & Kjaer) and custom software (Daniel Stolzberg). Calibrations were verified with a handheld spectrum analyzer (Bruel & Kjaer).

**Behavior**

Sinusoidal stimuli were presented to gerbils within the context of a standard aversive detection task. Gerbils were placed on controlled water access and trained to drink water continuously from a spout in the testing cage in the presence of any AM sound. Occasionally, the sound level held constant for 1.5 s, and a mild electrical shock would be delivered through the metal spout during the final 200 ms of these unmodulated noise trials. Gerbils learned to avoid the shock by withdrawing momentarily from the spout.

Stimulus presentation and timing of behavioral events were controlled using custom software (ePsych, Daniel Stolzberg), working in conjunction with an RZ6 multifunction processor (Tucker Davis Technologies; TDT). Recording sessions lasted 10–30 min, ending when the animal was satiated or after a sufficient number of trials were collected. All sessions analyzed showed good behavioral performance, with a mean d′ of 2.8 (d′ range 1.4–4.5; mean hit rate = 68%, mean false alarm rate = 2%).

**Electrophysiology**

*Surgery*

After gerbils were trained in the behavioral task, a silicon probe with either 16 or 64 recording sites was implanted in left auditory cortex (Neuronexus, model A4x4-4mm-200-200-1250-H16 and model Buzsáki64_5x12-H64LP_30mm). The probe was affixed to a custom-made manual microdrive, with a screw that allowed the electrode to be advanced parallel to the implantation plane. Probes were inserted at a 25° angle on a mediolateral axis such that advancement of the probe allowed sampling of multiple sites passing roughly tangentially through a cortical layer. Recording sites were spaced along four shanks (model A4x4-4 mm-200-200-1250-H16) or five shanks (model Buzsáki64_5x12-H64LP_30mm), with 200 $\mu$m between each shank, thus spanning the anterior/posterior tonotopic axis of AC. The surgery was performed under isofluorane anesthesia. Animals underwent one week of recovery before being placed back on controlled water access.

*Data acquisition and processing*

Electrophysiological data were acquired from freely-moving animals while they performed the aforementioned behavioral task using a wireless headstage and receiver (W16 or W64, Triangle Biosystems). Analog signals were amplified and digitized at a sampling frequency of 24,414 Hz and transmitted to a digital signal processor (TDT; 16-channel recordings: TB32 to RZ5; 64-channel recordings: PZ5 to RZ2) then sent to a PC for storage and postprocessing.

Electrophysiological data underwent common average referencing and were bandpass filtered at 300–5000 Hz. An artifact rejection procedure was performed to remove noisy portions of the signal, for instance from the electrical shock or extreme head movement. These trials were excluded from analyses. Open-source spike sorting packages were used to extract and cluster spike waveforms (16-channel recordings: UltraMegaSort 2000; 64-channel recordings: KiloSort), and manual sorting was performed on the output of the algorithm. All analyses were restricted to well-isolated single units (SUs). Unit quality was verified using several metrics, including separation in principal component space from other clusters, clear refractory periods, and waveform amplitudes above the noise floor throughout the recording sessions (Fig. 1*B*–*E*).

*Cell type assignment*

All high-quality single units that survived the spike sorting process were included in analyses; no response criteria were applied. Cells were labeled as regular spiking (RS) or narrow spiking (NS) according to spike width (Fig. 1*F*). The distribution of time from waveform trough to peak was bimodal, and a boundary was placed at 0.43 ms. This measure of categorization based on spike width was related to spontaneous firing rates (FRs): mean spontaneous FR of cells labeled as RS was 4.1 spikes per second, while mean baseline rate of NS cells was 13.3 spikes per second. In cortical neurons, these measures and categorizations are commonly thought to reflect excitatory and inhibitory neurons in cortex, respectively (Wilson et al., 1994; Barthó et al., 2004; Mesik et al., 2015; F. Liang et al., 2019).

*Histology*

The location of recording sites was confirmed to be in AC by histology. Subjects were intraperitoneally administered an overdose of sodium pentobarbital (150 mg/kg) and perfused (0.01 M PBS, 4% paraformaldehyde). The brain was extracted and postfixed in 4% paraformaldehyde. At the time of slicing, the brain was embedded in 3% agar and coronal sections of 50 $\mu$m were made on a vibratome (Leica). Sections were wet mounted onto gelatin-subbed slides and inspected under an upright microscope (Revolve Echo). Cytoarchitectonic features were used to find the closest matching plate in the gerbil brain atlas (Radtke-Schuller et al., 2016). While most brains were imaged using a fluorescent mounting solution containing DAPI (Vector Laboratories), one brain underwent a staining procedure for capturing images in bright field. In order to increase contrast of cytoarchitectonic landmarks for bright field imaging, we adapted a myelin staining procedure using Sudan Black (Ineichen et al., 2017). Figure 2 displays the sections that contain AC, marked in yellow, according to the Radtke-Schuller atlas. The site of the probe can be identified by perforations in the tissue and/or from damage at the dorsal surface, resulting from atrophy over the course of weeks to months of the chronic implantation. Perforations from the shanks of the probe are clearly visible and well aligned with primary AC (A1). Because it is possible that some units in the dataset came from secondary auditory regions, we refer to our recording site as AC.

**Neural analyses**

*Basic response properties*

Several traditional analyses were performed to assess the response properties of AC neurons to amplitude modulation stimuli. While spectral tuning was not directly analyzed, neurons were sampled along the entire tonotopic axis of AC, often simultaneously, as probe shanks spanned 600–800 $\mu$m anterior/posteriorly. Firing rate distributions were calculated from the average number of spikes emitted by each cell during the 1-s trials for each stimulus, and spontaneous rate from an epoch of silence at the beginning of each recording session. The resulting histograms were fit with log-normal distributions (Fig. 4*A*). The Kruskal–Wallis test was used to assess differences between distributions.

Each neuron's sinusoidal AM responses were characterized by traditional metrics. Firing rate (FR) was calculated for each periodic AM stimulus (1-s duration). A cell was considered significantly responsive if the FR distribution across stimuli passed the Kruskal–Wallis test ($p < 0.01$) and the stimulus with the highest response was significantly higher than the lowest (rank-sum, $p < 0.01$). The AM rate yielding the highest response was labeled that cell's best modulation frequency (rate BMF). To measure the proportion of significantly responsive cells for each periodic AM rate, the distribution of FR across trials was compared with the spontaneous FR distribution using the Wilcoxon rank-sum test at $p < 0.01$, Bonferroni corrected to 0.0002 (Fig. 4*B*). Collection of spontaneous firing occurred at the beginning of a recording session.

Temporal responsiveness was measured using the vector strength and the Rayleigh statistic ($p < 0.001$, Bonferroni-corrected; Fig. 4*C*). Temporal BMFs were defined for each cell as the periodic rate with the highest vector strength value, of those that were significantly synchronized. Mean response phases (Fig. 5*B*) were calculated excluding periods that began <250 ms from the onset of a trial to avoid any potential artifacts from stimulus transitions, and measuring the temporal dynamics of spiking responses during a "steady state" of AM.

The percentage of cells that showed significant adaptation or facilitation was calculated over the course of 1 s for each periodic AM rate.

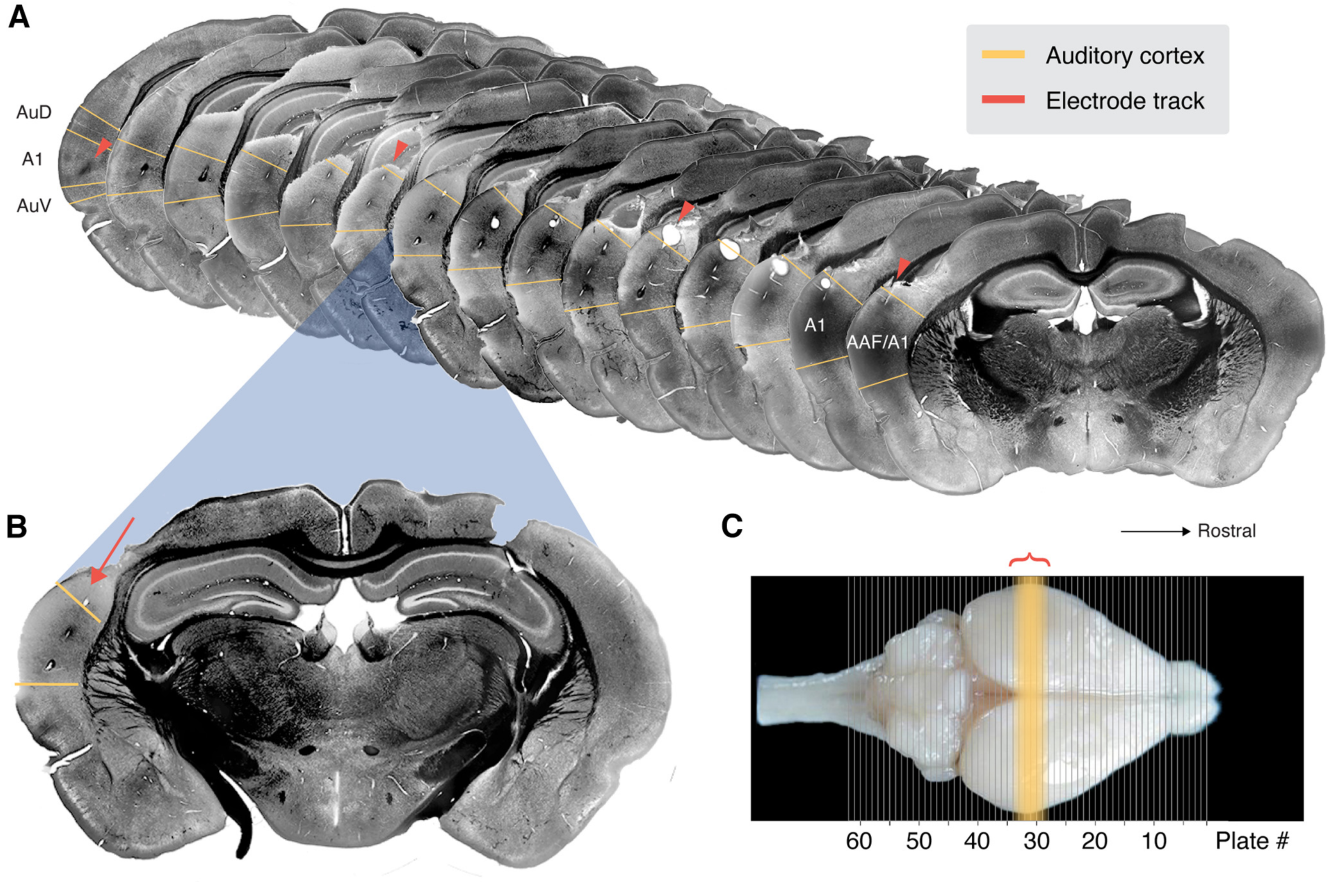

**Figure 2.** Histologic verification of probe location. ***A***, Coronal sections of the brain of one subject, myelin-stained with Sudan Black, spanning the full rostro-caudal range of core AC. Sections were matched to the coronal plates in the gerbil brain atlas (Radtke-Schuller et al., 2016), and the approximate boundaries of AC are marked in yellow. This range also corresponds to the sections that show perforations in the tissue from the shanks of the probe (orange arrows). ***B***, An enlarged section displays the probe location more clearly along with the cytoarchitecture of this section, which corresponds most closely with plate #30. ***C***, Dorsal view of an intact gerbil brain with plate locations overlayed (image from Radtke-Schuller et al., 2016). The yellow region marks those plates which contain AC according to the atlas, and the orange bracket marks the rostro-caudal span of visible tracks of our implanted probe.

Responses were labeled adapting if the number of spikes was significantly lower in the last 250 ms than the first, and vice versa for facilitating responses (Wilcoxon rank-sum $p < 0.01$, Bonferroni corrected). For the 2-Hz stimulus, spiking responses were compared between the first and last 500 ms.

*Population poststimulus time histograms (PSTHs)*

Analyses in Figures 5, 8–12 used responses from unique 500-ms segments that were extracted from the sinusoidal and vocoded speech stimuli. Cells were included in analyses if they had at least 12 trials for each segment. Activity was convolved with an exponential function ($\tau$ = 5 ms, binsize = 1 ms). For display in Figures 5 and 8, RS cells were split into five quantiles according to the mean of the maximum peak FR evoked by each stimulus. Thus, Q1 contains the cells exhibiting FR peaks in the highest 20% of the population, and cells in Q5 show little to no spiking modulation related to the stimulus segments. Within each group, cells are sorted from shortest to longest latency of the peak FR reached during stimulus three for sinusoidal AM (4 Hz) and stimulus 8 for vocoded speech ("Trees").

*Linear predictions of responses*

The linear generalizability of responses was compared across the stimulus classes, which had varying temporal complexity (Fig. 7). To do so, we adapted a forward model from the literature (David et al., 2009). Linear regression was used to estimate a temporal kernel for each cell from the exponentially-smoothed ($\tau$ = 5 ms) poststimulus time histogram (PSTH) created from 10 randomly drawn trials. To predict the cell's response to a different stimulus, the kernel was convolved with the stimulus amplitude trace. The quality of the prediction was quantified by Coincidence, or Pearson's correlation coefficient, between predicted and observed responses. Prediction quality was compared with the coincidence value resulting from a prediction based on the PSTH of 10 separate trials from the same stimulus type. The procedure was cross-validated by randomly drawing new trials on each of 100 iterations. Linearity was estimated for predictions of irregular sinusoidal responses based on periodic data, and for vocoded speech responses based on irregular sinusoidal data.

*Population activity surrounding envelope landmarks*

For the analysis presented in Figures 9 and 13, envelope landmarks, or features in the amplitude signal, were identified within the trial-averaged stimulus traces of Figures 5 and 8. First, the linear amplitude signal was transformed to a relative dB scale by taking the logarithm of the amplitude: dB = 20 · $\log_{10}(x)$ then subtracting the maximal value across all stimuli in the session. Local minima in the log-transformed amplitude signal were identified. Events were excluded if the ensuing local maximum was not at least 6 dB higher than the preceding minimum. peakDrv events were then defined as peaks in the derivative of the log-transformed amplitude signal, including only events that fell between valid minima and maxima events as described above. Events that followed a period of silence (i.e., intertrial intervals in vocoded speech sessions) were excluded from analyses, so all landmarks occurred within an ongoing acoustic stream. Thus, a peakDrv event represents a perceptually salient increase in amplitude that need not begin from silence.

Population PSTHs for each stimulus (binsize = 1 ms; exponentially-convolved, $\tau$ = 10 ms) were averaged across RS cells. A window of activity was extracted surrounding each valid peakDrv landmark. Sinusoidal AM results are presented separately for each stimulus, and vocoded speech results pool all events extracted across stimuli (22 total for vocoded speech). Results for vocoded speech show the mean ± SD across events.

*Classification of temporal structure by single units*

A machine learning classification approach was used to quantify the ability of each cell's spiking activity to discriminate between modulation rates (Figs. 10, 13). Eight envelope segments (those from Fig. 5 for sinusoidal AM and Fig. 8 for vocoded speech) were extracted from the full stimulus set and used in an eight-way classification task. Broadly, the classification procedure measured how reliably single-trial spiking responses could correctly discriminate stimulus tokens.

For each cell, one iteration of the classification procedure occurred as follows. A total of 16 trials of each stimulus token were randomly drawn and used for training the classifier, and one additional trial of each was set aside for testing. Thus, at least 17 trials of each stimulus were required for units to be included in this analysis. The number of trials was determined as the minimum number of trials needed to achieve reliable templates and avoid overfitting the training data. Commonly, classification algorithms are applied to spiking activity directly. Instead, we implemented a preprocessing step, which imparted a few benefits. First, it made the procedure less sensitive to slight offsets of phases between AM stimulus segments, or from trial to trial. Additionally, this preprocessing step mitigates the risk of overfitting introduced by increasing dimensionality of input data to the classifier. This mitigation measure becomes important for population decoding analyses (described in the following section), when evaluating classification accuracy as additional cells are added to an ensemble.

To create the training data for the classifier, for each stimulus, the dot product was calculated between each of the 16 trials and the averaged activity of the other 15 trials. Thus, each of the eight stimuli was represented by a distribution of projection values. As in prior analyses, spiking activity was binned at 1 ms and exponentially convolved ($\tau = 5$ ms) before these projection values were calculated. These data were fed into a support vector machine (SVM) with a linear kernel (templateSVM and fitcecoc functions in MATLAB). The algorithm learned the discriminant functions that best separated stimulus classes defined in the training data. We then calculated the dot product between the held-out test trial and the trial-averaged response template for each stimulus. If the single-trial spiking response to a given stimulus consistently differed in firing rate or pattern from the other stimuli, classification was reliable. This procedure was repeated for 500 iterations with 17 new randomly-drawn trials, and was performed for each cell that had a sufficient number of trials of each stimulus token (≥17). Classifier results were obtained in the form of a confusion matrix (Fig. 10*A*,*C*), then transformed into a signal detection theory metric, d′ (Green and Swets, 1966). The discriminability of each stimulus was calculated from the confusion matrix by the following equation:

$$d^{'}\text{stim} = z(\textit{hit rate}) - z(\textit{false alarm rate}).$$

The classification accuracy assigned to a cell was the mean of d′ across stimuli. Overall, a consistent pattern of spiking emitted in response to a stimulus, unique from the spiking patterns evoked by other stimuli, results in good classification and high d′ values. All main findings were confirmed using alternative approaches to classification (e.g., SVM input data consisting of spiking vectors instead of projection values, different binning and convolution parameters).

*Classification by neural ensembles*

The spiking of individual cells is variable trial-to-trial, particularly in relation to sound envelope. Therefore, we investigated strategies for population decoding, to assess how a downstream neuron may be able to extract reliable information about the envelope signal based on the activity in AC during a single trial. The ability of a population of cells to discriminate between AM stimuli was quantified using a similar SVM classification procedure to that described above, and two methods of pooling information across cells were compared (Fig. 12). For one approach, trial-template dot products were calculated as above for each cell individually, and all projection values included as input to the SVM algorithm. This method allowed the classifier access to information from each cell individually and is referred to as "Independent" pooling (plotted in blue). In the other method, referred to as "Summed" pooling, trial-template projections were calculated from spiking activity summed across cells (plotted in yellow). This approach eliminates any information that might be carried by cells' individual firing patterns, and it assesses classification accuracy based on the collective temporal pattern of activity across the population. Control analyses were run to confirm that the difference in dimensionality did not affect classification results. For these analyses, RS cells were pooled across sessions and animals. Results are reported as d′ values capped at four and reflect the average performance of 500–1500 iterations of the classification procedure.

The same approaches were applied to quantify classification accuracy of smaller ensembles of cells. Maximally informative populations can be approximated by selecting units in descending order of the information they carry individually (Ince et al., 2013). Beginning with the best SU, we gradually expanded the pool size and quantified classifier performance for each ensemble for both Independent and Summed population pooling methods.

## Results

In order to characterize the population-level representation of sound envelope in AC, we recorded extracellular single unit activity in core AC of freely-moving gerbils in response to several types of amplitude modulated noise. We first analyzed encoding of sinusoidal AM in the cortical population, looking at single units and at an aggregated population signal. We then extended these observations to describe cortical coding of one-channel vocoded speech. We used the observed encoding patterns to make predictions about decoding strategies, comparing between methods with the goal of identifying envelope cues to support segmentation of a continuous acoustic stream.

The sinusoidal AM stimuli consisted of a continuous stream of sinusoidally-modulated noise, which included both periodic and irregular intervals built from AM rates in the 2 to 32 Hz range (Fig. 1*A*; see Materials and Methods). The two irregular stimuli were constructed by permuting the sequence of, then concatenating, two full periods of each periodic AM rate. Periodic stimuli were 1000 ms in duration, and irregular stimuli were 1938 ms. Neural activity was recorded while animals were in an alert, engaged state. Subjects performed a straightforward perceptual task, which served to limit variability from changes in head position and behavioral state (details in Materials and Methods).

Extracellular activity was recorded in core auditory cortex (AC) with 16-channel or 64-channel silicon probes (Fig. 1*B*). Single units were sorted offline and confirmed to be well isolated based on several quality metrics (Fig. 1*C*–*E*). A bimodal distribution of spike widths allowed us to separate cells into regular spiking (RS) and narrow spiking (NS) subpopulations (Fig. 1*F*), a distinction that is correlated with excitatory and inhibitory cell types (Mesik et al., 2015; F. Liang et al., 2019). This report analyzes 277 single units, 205 RS and 72 NS, and their responses to sinusoidal AM. We also collected responses to vocoded speech stimuli, introduced in more detail later, and this dataset consisted of 130 single units (100 RS and 30 NS). We neither searched for strong responses nor applied any *post hoc* responsiveness criteria to exclude cells from the dataset, in effort to collect a broad and unbiased sample of the entire AC population. After recording sessions were completed, histology was performed to confirm that the probes were located in AC (Fig. 2; see Materials and Methods).

**Summary of sinusoidal AM response properties**

Figure 3 displays the spiking activity of three example cells during each stimulus in the sinusoidal AM set. The spike width of

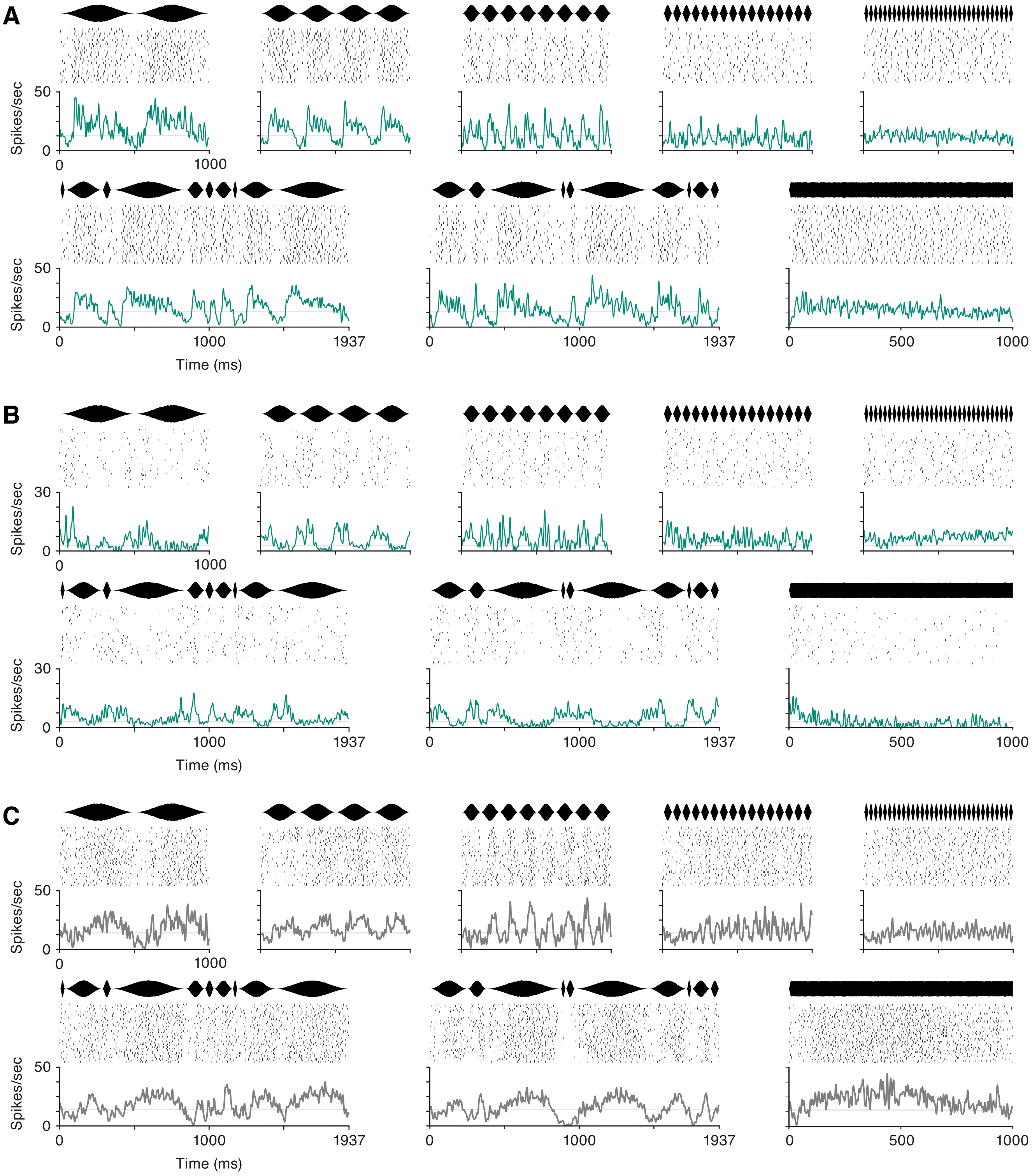

**Figure 3.** Example SU responses to sinusoidal AM stimuli. ***A***, Responses of an RS cell are shown for periodic stimuli (top: 2, 4, 8, 16, 32 Hz), both irregular stimuli, and unmodulated noise. At the top of each panel is the average stimulus envelope. Below, Raster plot shows spike times during the first 20 trials. Trial-averaged activity is shown at the bottom. ***B***, Responses of another RS cell displayed using the same conventions. ***C***, Responses of an NS cell to each sinusoidal AM stimulus. Each cell's mean spontaneous firing rate during silence is marked with a gray line in each panel.

the first cell classified it as a RS unit (Fig. 3*A*). Its firing rate synchronized to several AM stimuli, with firing peaking early in the cycle yet sustained throughout the period. The middle cell, also an RS unit, synchronized better to slower AM rates and showed a different preferred response phase (Fig. 3*B*). This cell's spiking peaked near the onsets and offsets of 2- and 4-Hz periods. The last cell had a narrow spike width, categorizing it as NS (Fig. 3*C*). Its firing rate modulated with the AM cycles, closely echoing the shape of the noise.

To gain a sense of the diversity of response properties both within and across RS and NS cell types, responses of all recorded cells were quantified according to several traditional response

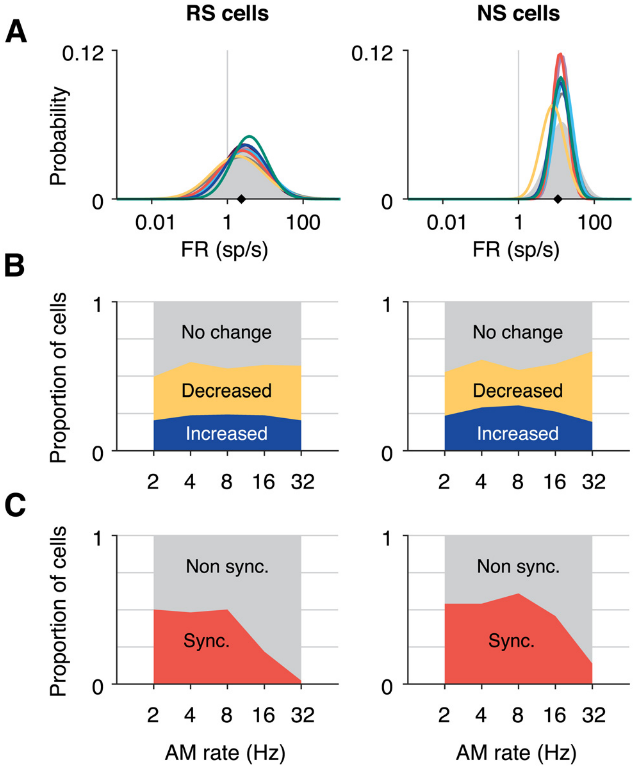

**Figure 4.** Tuning characteristics to sinusoidal AM stimuli. ***A***, Probability distributions of firing rates (FRs) during each stimulus (colors) and silence (gray area), for RS cells (left) and NS cells (right). For each group, data were fit with a log-normal distribution. NS FRs are higher than RS cells (ANOVA $p = 1.31e^{-134}$; overall medians for each group are denoted by black diamonds: RS cells = 2.35 spikes per second, NS cells = 11.08 spikes per second). Within RS cells, 32 and 16 Hz (yellow and dark gray lines) yielded firing rates significantly lower than one irregular stimulus (green line; Kruskal–Wallis $p = 0.007$, Bonferroni *post hoc* correction). For NS cells, none of the stimulus FR distributions significantly differed (Kruskal–Wallis, $p = 0.073$). ***B***, Proportion of RS and NS populations responsive to each periodic AM stimulus as measured by change in average FR. The proportion of units in which FR significantly increased from spontaneous rate is shown by the blue area for each AM rate, and the proportion of units with decreased FR from spontaneous is shown in yellow (significance determined by Wilcoxon rank-sum, $p < 0.01$, Bonferroni corrected). ***C***, Proportion of RS and NS cells that were significantly synchronized is shown in orange for each AM rate (significance determined by Rayleigh statistic, $p < 0.01$, Bonferroni corrected).

metrics. Average firing rates (FR) during each stimulus were roughly log-normally distributed and were consistent across periodic AM rate and irregular stimuli (Fig. 4*A*). The distribution of spontaneous rates, measured during silence, were also remarkably similar (filled gray area). Firing rates of NS cells were greater than those of RS cells (ANOVA $p = 1.31e^{-134}$; black diamonds in Fig. 4*A* and 4*B* mark the median FR across all stimuli: RS cells = 2.35 spikes per second, NS cells = 11.08 spikes per second). For RS cells, FR distributions for 16 and 32 Hz (dark gray and yellow lines, respectively) were significantly lower than the second irregular stimulus (green line; Kruskal–Wallis $p = 0.007$). Within NS cells none of the distributions significantly differed from one another. These findings are in agreement with previous studies in rodent models (Hromádka et al., 2008; Hoglen et al., 2018).

The stability of FR distributions might suggest that the stimulus had little effect on neural activity in the population. To unpack these population FR distributions, we measured each cell's direction of change from spontaneous activity for each periodic AM rate (Fig. 4*B*). Approximately 25% of cells increased their firing above baseline (blue), ~25% were suppressed (yellow), and 50% showed no change in the time-averaged FR compared with silence (gray). These proportions were similar in RS and NS cells, and they demonstrate that excitatory and inhibitory responses did occur in many cells but were balanced across the population such that average FRs did not change. AM rate tuning was also measured. The distribution of best modulation frequency (BMF) as measured by FR was roughly uniform in this range of AM rates (data not shown). This heterogeneity of AM rate tuning is in agreement with previous literature (Schreiner and Urbas, 1988; Eggermont, 1998; L Liang et al., 2002; Joris et al., 2004; Malone et al., 2007; Zhou and Wang, 2010; Yin et al., 2011; Hoglen et al., 2018).

Neurons also encode AM with phasic modulations of spiking over time, synchronized to the stimulus. To gauge the prevalence of temporal responses in the population, we quantified the percentage of cells with significantly synchronized activity according to the Rayleigh statistic (see Materials and Methods) during each periodic AM rate (Fig. 4*C*). Approximately 50% of RS and NS cells were phasically modulated by AM rates of 8 Hz and below (orange). In line with previous descriptions of cortical phase locking limits in rodents, synchronization fell off at 16 and 32 Hz. NS cells were more likely to continue phase locking at higher AM rates.

When temporal and rate response metrics are considered together, 60–70% of RS cells and 70–80% of NS cells were considered responsive by at least one of these two measures for each periodic AM rate (data not shown). These high proportions of responsive cells stand in contrast to the overlapping FR distributions of Figure 4*A*. Specifically, the AM stimulus modulated the activity of a majority of cells in the population, but changes in firing rate were balanced across time and across cells in a way that maintained an overall equilibrium.

## Responses are stereotyped at the beginning of slow modulation events

The analyses presented above conform with the literature in demonstrating that AC cells show heterogeneous AM tuning as measured by firing rate and synchronization. However, these traditional response metrics fail to describe the alignment of spiking with the dynamic changes in amplitude that define AM stimuli. Mean phase of firing has been evaluated within cell as stimulus parameters are varied, but the temporal structure of activity in relation to the stimulus remains unexamined. If the timing of synchronized responses is heterogeneous, like the response metrics above, mean phase would be distributed evenly throughout the AM stimulus, tiling each modulation cycle. Alternatively, certain envelope features could be overrepresented.

First, responses in the population collectively were visualized by plotting the average activity of each cell during each of eight 500-ms tokens extracted from the sinusoidal AM stimulus set (Fig. 5*A*). RS cells were grouped into five quantiles of equal size according to their peak firing rates (see Materials and Methods). Cells within each RS group were sorted by the time of the peak response during the 4-Hz token, such that cell identity remains constant as a row across panels. NS cells were sorted by time of the firing minimum during 4 Hz.

A large portion of neurons displayed temporal structure related to the stimuli, in accordance with the summary metrics of Figure 4*C*. Qualitatively, synchronized responses to 2 Hz appeared to be more prevalent near the beginning of a modulation period, while, in contrast, responses to faster modulations appeared to distribute through the modulation cycles.

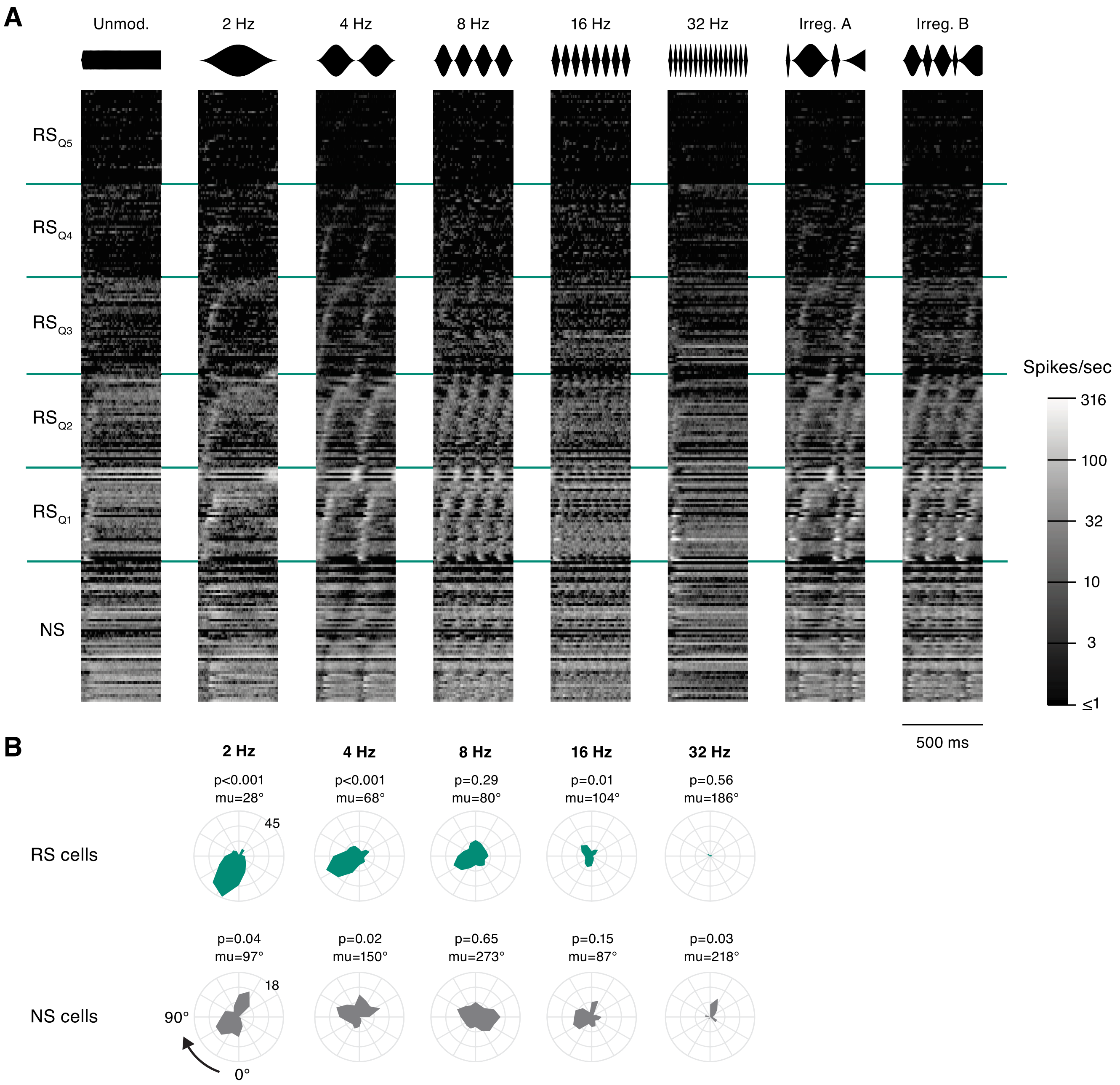

**Figure 5.** Response timing is biased toward onsets of modulation cycles. ***A***, The average response of each neuron during eight segments of sinusoidal AM is plotted on a color scale, where lighter gray indicates a higher FR. NS cells are grouped at the bottom, and RS cells are split into five quantiles according to maximum peak FR. Cells within each RS group are sorted by the latency of peak FR during the 4-Hz stimulus, and NS cells are sorted by the time of minimum FR during 4 Hz. The identity of a cell is maintained in a row across all stimuli. Stimulus waveforms are illustrated above each column. ***B***, For each periodic AM rate, the mean phase distribution is shown for all synchronized cells, RS above and NS below. Distributions are shown on polar plots that represent one modulation cycle. The trough, phase = 0, is at the bottom and the period proceeds clockwise. The numbers positioned at 225° label the limits of the radial axis, which corresponds to number of cells. The *p*-value above each plot denotes significance level from the Rayleigh test of nonuniformity. The mean preferred phase across all synchronized cells (mu) is listed above each plot.

To quantify the timing of responses, we plotted the mean phase distributions for each periodic stimulus (Fig. 5*B*). For each cell with a significantly synchronized response to that AM rate, its mean phase was calculated, quantifying when during an AM period spikes were likely to occur. The distributions of cells' preferred phases are presented as polar histograms, in which the onset (trough) of a period is at the bottom and the polar axis proceeds clockwise. The length of the polygon on the radial axis illustrates the number of cells with that mean phase. While mean phases were heterogeneous in many cases, responses were stereotyped within RS cells during 2- and 4-Hz stimuli. The Rayleigh test for nonuniformity confirmed that these distributions were significantly skewed (2 Hz: $p = 1.44e^{-14}$; 4 Hz: $p = 6.78e^{-8}$), collecting around 28° for the 2-Hz stimulus and 68° for the 4-Hz stimulus. As each period was part of a continuous stream of amplitude-modulated noise, this bias toward cycle onset is not the result of sound onset in the classical sense, i.e., when preceded by silence. In fact, to ensure our phase measurements came from "steady-state" AM, the first 250 ms following transitions between different AM rates was excluded from mean phase calculations. Thus, responses of the AC population were biased toward the onsets of periods during continuous low-frequency modulation.

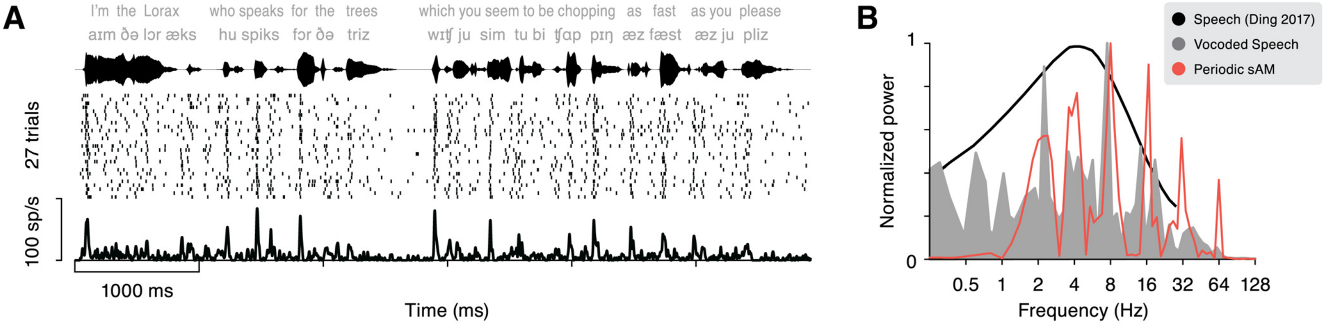

**Figure 6.** Vocoded speech stimuli. ***A***, The example response of an RS cell is plotted for one of the six vocoded speech stimuli. The stimulus waveform is shown on top, and the raster and histogram are plotted below. ***B***, Power spectra showing the relative energy across modulation frequencies, for comparison across stimulus sets. The power spectrum calculated from periodic sinusoidal AM stimuli is plotted in orange, and that computed from vocoded speech stimuli is shown in gray. The black line illustrates the modulation power spectrum of a large database of recorded speech (adapted from Figure 3*A* in Ding et al., 2017).

### Nonlinearity in envelope responses based on local temporal dynamics

The sinusoidal stimuli in this report focus on the slow modulation frequencies that are known to be important for the processing of human speech and other animal communication sounds. However, it is not clear that observations based on simple, periodic stimuli directly translate to more complex envelopes such as those of human speech.

To compare the sinusoidal AM data to a more natural signal, neural activity in AC was also collected in response to noise modulated by the envelopes of natural speech. 130 single units were recorded in this condition: 114 RS and 16 NS cells. Of these, 41 cells were also recorded during sinusoidal AM. An example cell's response to one of the six vocoded speech stimuli is shown in Figure 6*A*.

Sinusoidal (orange line) and speech (filled gray area) stimuli showed overlapping modulation power spectra (Fig. 6*B*), and roughly corresponded to the power spectrum calculated from a large database of natural human speech (black line; adapted from Ding et al., 2017). However, the local temporal properties of speech envelopes are more complex than sinusoidal AM. For instance, natural sounds are rarely symmetrical: the shape of the amplitude ramp and decay is not directly determined by the repetition rate. How well do responses to periodic, sinusoidal AM stimuli predict neural responses to other envelopes? If neurons track sound amplitude in real time, the response to any envelope should be linearly predictable from that of periodic AM stimuli. On the other hand, if responses differ from the linear prediction, it exposes a nonlinear relationship between the stimulus and spiking activity.

To probe the linearity of responses, we first assessed the impact of periodicity in sinusoidal AM, adapting a common forward model to predict FR over time (David et al., 2009). Here, the procedure gauged the ability of responses to periodic stimuli to predict responses to irregular stimuli. For each cell, a linear kernel was learned from a PSTH of its responses to all periodic stimuli (see Materials and Methods for more details). The resulting kernel was used to predict the cell's response to the irregular AM stimuli (Fig. 7*A*, orange). The quality of the prediction was measured by the coincidence (C) of the predicted response, based on 10 randomly-drawn trials, and the observed response, measured from a held-out set of 10 trials. To stabilize coincidence values, this procedure was cross-validated 100 times per cell. These periodic-to-irregular coincidence values were compared, within-cell, to coincidence values derived from predictions of irregular responses based on kernels learned from irregular data (Fig. 7*A*, blue). The resulting coincidence values are plotted in Figure 7*B* by cell type (RS: filled green circles, NS: open circles; example cell from Fig. 7*A* is RS and filled in black). Periodic stimuli produced similar coincidence values as predictions created from the held out irregular data (Pearson's $r = 0.98$, $p = 1.66e^{-194}$). Predictions from irregular data were slightly but significantly better ($\Delta C_{\text{Pdc-Irr}} = -0.006$, $p = 1.59e^{-4}$ Wilcoxon sign-rank).

One possible contribution to the difference in linear prediction quality could be the small proportion of cells that displayed adaptation or facilitation during a periodic stimulus of 1-s duration. For AM rates of 2–8 Hz, <3% of cells, RS or NS, demonstrated any significant change in firing rate from the beginning to the end of the stimulus (rank-sum $p < 0.05$, Bonferroni corrected). In response to 16 Hz, most RS cells' FRs were still constant, while slightly more NS cells displayed either adaptation or facilitation. At 32 Hz, many more cells demonstrated changes in FR: 7% (RS) and 6% (NS) showed a significant increase in firing and 9% (RS) and 31% (NS) of cells displayed adaptation. Consistent with prior results in awake squirrel monkeys (Malone et al., 2015), this observation could contribute to a difference in linearly predicted responses because the adaptive processes that occur during sustained periodic modulations would less likely be engaged during irregular AM stimuli.

Next, the ability of irregular sinusoidal data to predict responses to vocoded speech was assessed for the 41 cells that had a sufficient number of trials in both stimulus conditions (Fig. 7*C*,*D*). While the response of a cell to sinusoidal AM did provide information about vocoded speech responses (Pearson's $r = 0.74$, $p = 3.18e^{-8}$), speech-to-speech predictions yielded higher coherence values on average ($\Delta C_{\text{Sin-Speech}} = -0.06$, $p = 3.91e^{-7}$ Wilcoxon sign-rank). The predictive power for AM to speech stimuli was an order of magnitude smaller than comparing periodic to irregular AM, which suggests that drawing conclusions about speech representations based on AM data should be done carefully. Overall, these analyses demonstrated that the envelope representation was somewhat robust across different types of modulations. However, just as AC cells demonstrate spectral (Sadagopan and Wang, 2009) and spectrotemporal (David et al., 2009) nonlinearities, temporal nonlinearities may exist independently, as well. Because irregular AM and vocoded speech stimuli contained a similar distribution of energy across modulation frequencies on average, the nonlinearities were likely driven by real-time envelope dynamics. This result emphasizes the notion that the AC

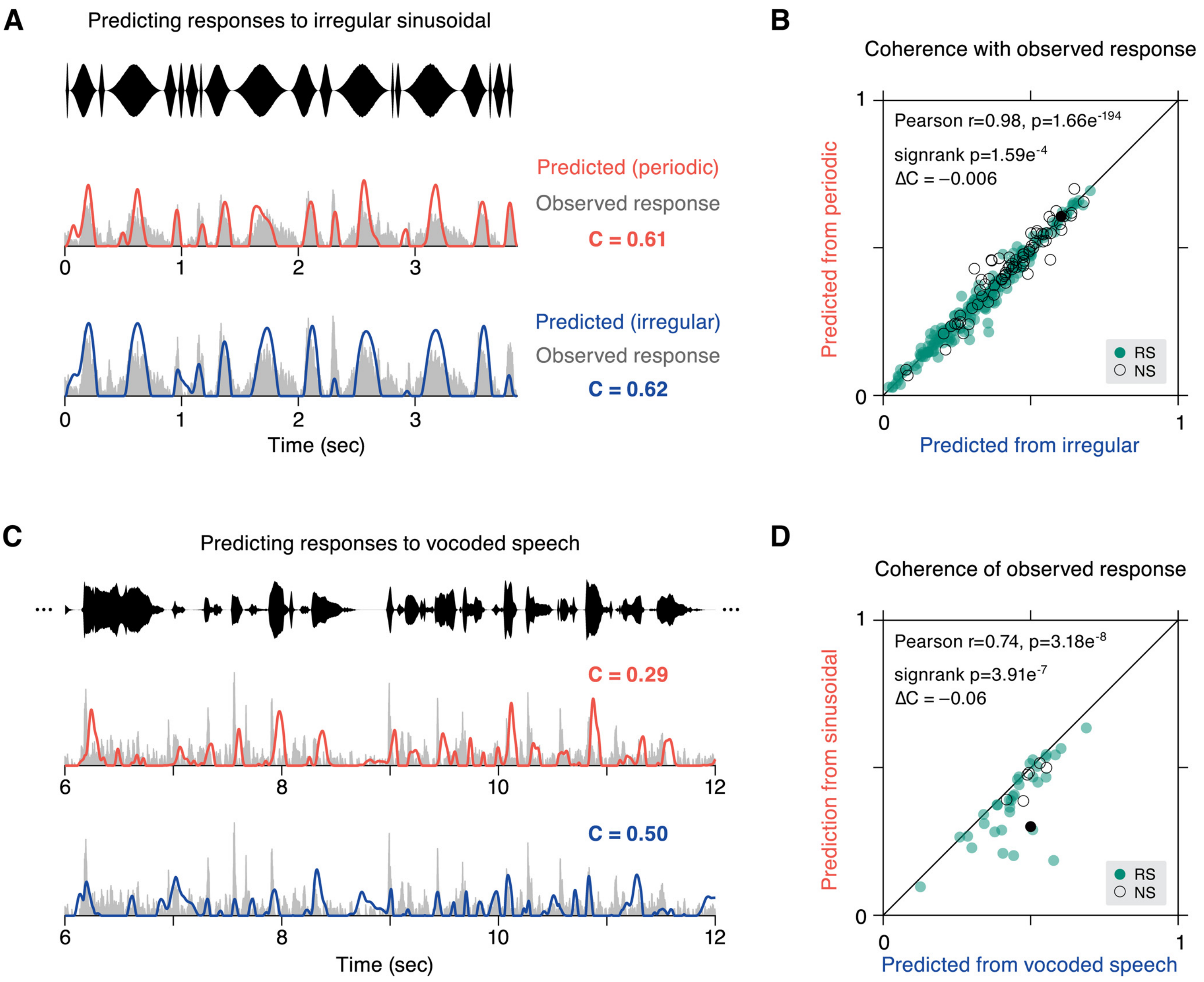

**Figure 7.** Nonlinearity in responses from local features of modulation stimuli. ***A***, Two predictions for the response to irregular modulation were created, one from a linear kernel generated from periodic data, and the other from a kernel based on held out trials of irregular data. The quality of each prediction was quantified by a coincidence value (C). Predicted (orange) and observed (gray) responses are plotted for an example cell for periodic training data and irregular training data (blue). ***B***, Coincidence values for each cell of observed responses with responses of the same stimulus (irregular data, abscissa) are shown against coincidence values obtained from the kernel constructed from the opposite stimulus type (periodic data, ordinate). Overall, linear predictions are similar for either context (Pearson's $r = 0.98$, $p = 1.66\text{e}^{-194}$), although same context predictions are slightly higher ($\Delta C_{\text{Pdc-Irr}} = -0.006$, $p = 1.59\text{e}^{-4}$ sign-rank). ***C***, The ability of sinusoidal data to predict vocoded speech responses was compared with predictions generated from separate trials of vocoded speech. Predicted activity for an example cell is shown for the opposite stimulus type (irregular sinusoidal, orange) and same stimulus (vocoded speech, blue). ***D***, Coincidence values are plotted for all cells that were recorded in both stimulus conditions ($N = 41$), with predictions from the same stimulus type (vocoded speech, abscissa) and the opposite stimulus type (irregular sinusoidal, ordinate). Again, coincidence values are strongly correlated across stimulus types (Pearson's $r = 0.74$, $p = 3.18\text{e}^{-8}$), but same context prediction is better ($\Delta C_{\text{Sin-Speech}} = -0.06$, $p = 3.91\text{e}^{-7}$ sign-rank). Colors in panels ***B***, ***D***: RS cells are filled green and NS are outlined black. Example cells from ***A***, ***C*** are highlighted as filled black circles.

representation is not simply an analog reproduction of the sound envelope, but a nonlinear, temporal transformation of the sensory signal.

### Coherent population responses from envelope landmarks

If cortical neurons do not linearly encode the sound amplitude, what features of the envelope drive responses? Figure 8 shows the average activity of each cell during eight 500-ms unique segments extracted from the full vocoded speech stimuli. Like sinusoids, speech-shaped modulations also evoked temporally rich patterns of activity in the population of AC neurons. While the shapes and timing of responses are heterogeneous across cells, it qualitatively appears that the onsets of syllables are often correlated to increased spiking across many cells in the population.

The temporal heterogeneity of responses to irregularly shaped waveforms cannot be quantified with the standard circular statistics used for periodic stimuli (as in Fig. 5*B*). Instead, we borrowed from an analysis of electrocorticographical (ECoG) data recorded during speech processing in humans (Oganian and Chang, 2019). This analysis inspects the pattern of neural activity surrounding a particular feature of the speech envelope (e.g., local peaks or local minima in the amplitude signal). Specifically, Oganian and Chang (2019) demonstrated that a sharp increase in high $\gamma$ activity follows peaks in the derivative of the speech envelope. Further, they found that these acoustic landmarks correspond to a linguistic landmark: syllable nuclei. Oscillatory activity in the high $\gamma$ range is thought to reflect an aggregate of local neuronal firing, making this observation congruent with the biased phase preferences exhibited in Figure 5*B*. Thus, we adapted the analysis from Oganian and Chang to ask whether peaks in the derivative of amplitude (peakDrv) also corresponded to a coherent increase in firing rate of individual AC cells in gerbils.

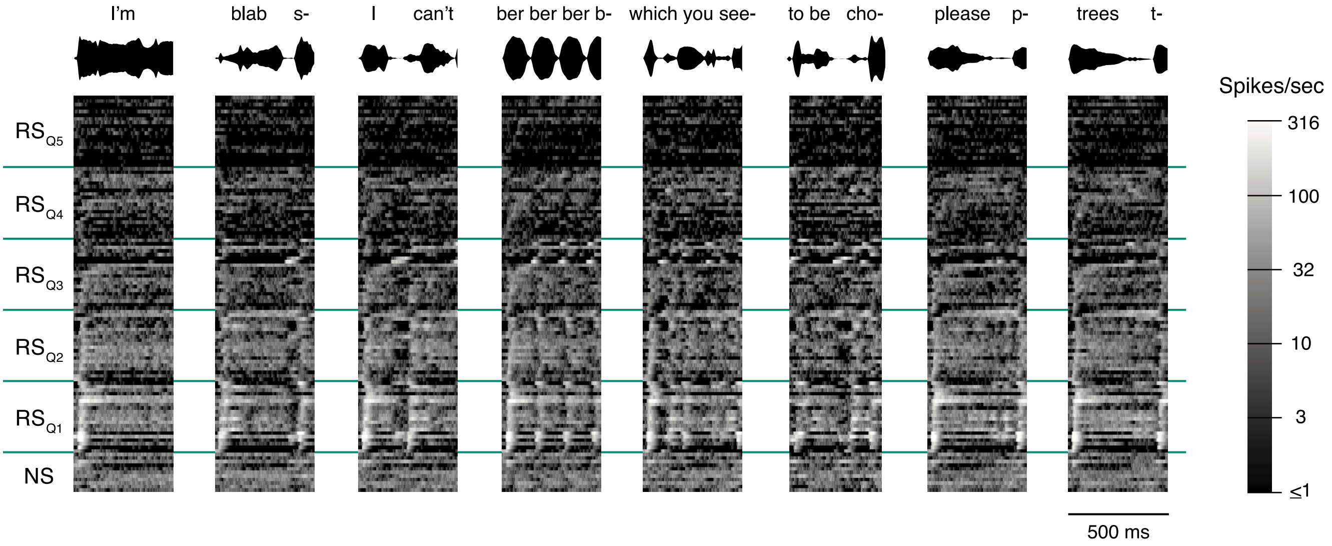

**Figure 8.** Response patterns in the cortical population during vocoded speech. Average responses for all cells with at least 12 trials to segments drawn from speech-derived AM stimuli, shown with the same conventions as Figure 5. Stimulus waveforms are illustrated above each column. Each cell's trial-averaged response profile is plotted with lighter gray indicating a higher FR. NS cells are grouped at the bottom, and RS cells are split into five quantiles according to peak FR. Each group is sorted by the latency of the peak FR (for NS cells: minimum FR) reached during the stimulus on the far right: "Trees." The identity of a cell is maintained across a row.

Landmarks were identified in the envelope signal, defined as local peaks in the derivative of the logarithmic transformation of the sound envelope (peakDrv; Fig. 9*A*, orange dots). Extracting peakDrv landmarks from speech envelopes was less straightforward than from sinusoidal AM because natural stimuli contain many local maxima and minima (Fig. 9*B*). Thus, we set a threshold, requiring the sound to double in sound pressure amplitude (+6 dB) to constitute a valid peakDrv event. To apply this threshold, we transformed the linear stimulus amplitude to a decibel (dB) scale, to reflect the SPL profile of the modulation (Fig. 9*A*, right). Note that, as this rescaling is nonlinear, it means that peakDrv events (orange dots) occur closely in time to envelope minima (yellow dots), or roughly the onsets of modulation cycles (Fig. 9*A*; e.g., separated by 8 ms in the 4-Hz stimulus). See Materials and Methods for full description of event identification. Ultimately, peakDrv events represent perceptually salient increases of amplitude in an ongoing sound stream.

After peakDrv landmarks within stimuli were identified, spiking activity was summed across all RS cells and the evoked response was calculated by averaging the population activity surrounding each peakDrv event. The schematic in Figure 9*B* and 9*C* illustrates this process for one vocoded speech stimulus. When summed activity of RS cells was aligned to peakDrv events in the sinusoidal AM stimulus, temporally synchronized firing was prominent (Fig. 9*D*). Population activity peaked with latencies between 30–90 ms, and the phasic response was larger in magnitude and spread over a broader time period for slower modulation frequencies. peakDrv events during irregular stimuli evoked a similar phasic increase in population firing rate, with an attenuated peak because of averaging responses across the range of AM periods (data not shown). The same analysis was performed for landmarks during the complex envelopes of vocoded speech. When population activity was aligned to peakDrv events in vocoded speech, neural activity showed a prominent peak around 40 ms following these landmarks (Fig. 9*E*).

When activity was averaged around local peaks in the amplitude envelope of vocoded speech stimuli, the resulting population trace showed a shallower, broader peak of activity centered at 3-ms latency (data not shown). The fact that the firing rate began increasing long before amplitude peaks occurred implies that this landmark is not likely responsible for evoking coordinated activity in the AC population. When local minima in the envelope signal were used as landmarks for this analysis, evoked activity traces looked nearly identical to those yielded from peakDrv events (data not shown).

Taken together, the results confirm that amplitude edge onsets are overrepresented at the population level. While the mean phase histograms in Figure 5*B* suggested that many cells are triggered near onset events, the present analysis shows that this encoding bias results in a pattern of activity that is visible in the global population signal. This brief coherence in the population could contribute to the global signals recorded by EEG and ECoG studies. As observed for rhythmic tracking in human AC, the effect observed here is strongest for modulation frequencies at or below the natural rate of syllables.

### Decoding envelopes from individual neurons is unreliable

The presence of a coherent, global signal predicts that prominent amplitude edges could be decoded by a strategy that ignores the heterogeneity and tuning of individual cells. In other words, sampling the aggregate level of activity across the AC population could be sufficient to parse an acoustic stream into behaviorally-meaningful segments.

While the preceding analyses examined patterns in trial-averaged activity, the perceptual and behavioral effects of a sound stimulus in natural scenarios are driven by the collection of noisy spike trains emitted by cells. To investigate how cues for envelope parsing could be decoded from single-trial population activity in AC, we performed several classification analyses. The results presented above describe a representation that could be useful for segmenting a continuous acoustic stream into smaller windows for subsequent processing. The parsing process must occur in real time, and acoustically, ground truth is ambiguous as boundaries are defined perceptually. Thus, instead of arbitrarily defining boundaries to train the classifiers to detect, classifiers were trained to categorize neural responses to short envelope segments which differed only in their temporal patterns of modulation.

Before decoding from the collective population, the classification performance of each cell was first assessed individually. A

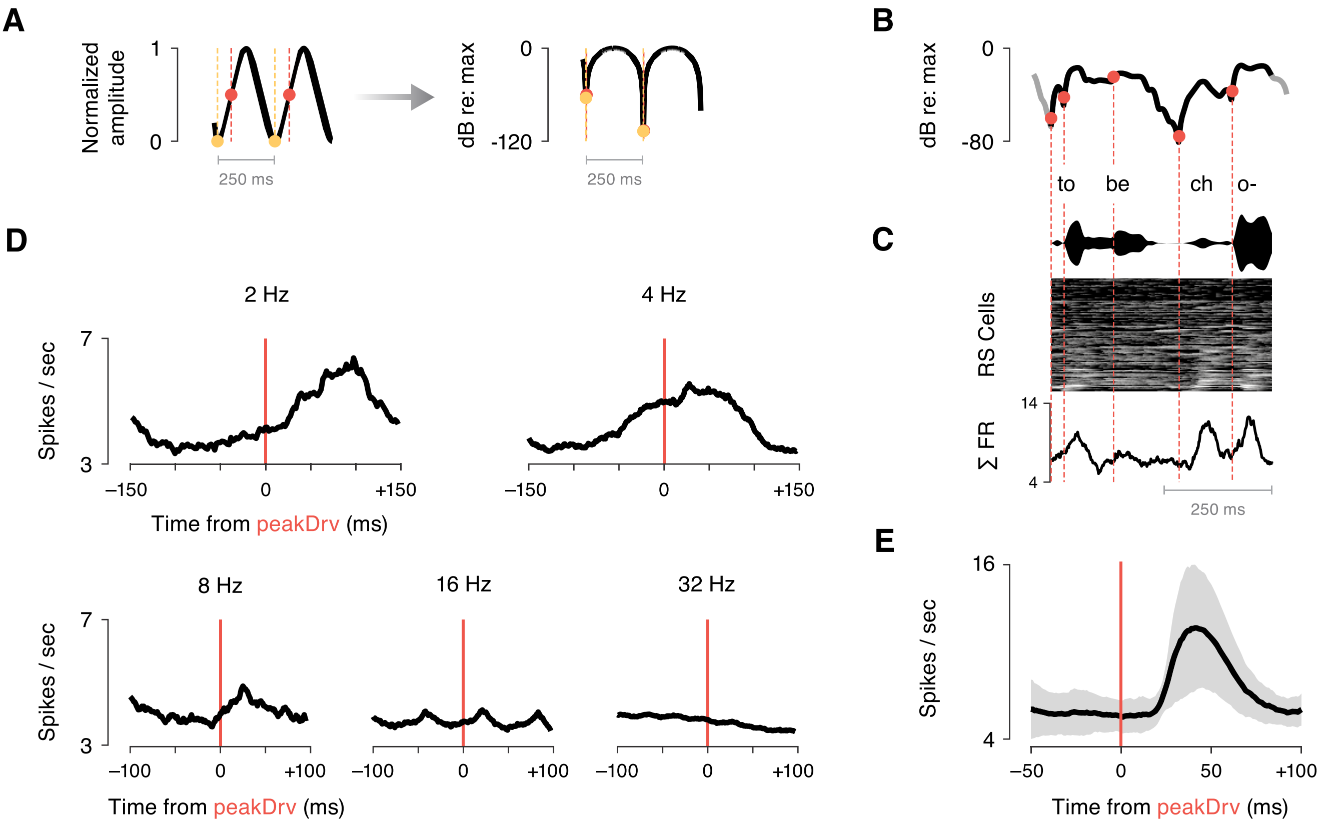

**Figure 9.** Coherent population responses follow peak derivative events during slow modulations. We measured evoked responses in the global average firing rate of RS cells, reasoning that if a given amplitude feature evokes spiking in many cells, this coherent response would be reflected in the mean population activity. ***A***, A schematic illustrating the identification of peakDrv events (orange circles) in sinusoidal AM stimuli. The left panel marks these events on a linear amplitude scale. For this analysis, peakDrv landmarks were identified within the logarithmically-transformed amplitude signal (right panel), corresponding to perceptual space. Note that, when the amplitude is expressed on a dB scale, peakDrv events coincide with the onsets (minima) of sinusoidal amplitude cycles (yellow circles). ***B***, peakDrv events were identified within the log-transformed envelopes of vocoded speech stimuli. To focus on perceptually relevant amplitude modulations, we restricted analyses to peakDrv events that occurred between a local minimum followed by a local maximum at least 6 dB higher. The schematic in panel ***B*** illustrates the peakDrv events identified within an example speech segment. ***C***, The stimulus envelope and corresponding RS population activity are depicted using the conventions of Figure 8. The trace at the bottom shows the mean activity across all RS cells for this stimulus segment. ***D***, Mean population activity surrounding peakDrv landmarks was averaged across all events, plotted for each periodic modulation rate. Slower rates evoke stronger phasic activity in the population, and the short, positive latencies suggest that peakDrv could have a causal relationship to coherent activity in the population. ***E***, Evoked population activity is plotted after averaging across peakDrv events in all vocoded speech stimuli (mean ± SD). This stimulus feature evoked a strong, phasic response in the aggregate population with ∼45-ms latency. The rising edge of the evoked response is sharp, as would be expected if the causal stimulus landmark were aligned in time.

linear support vector machine (SVM) was trained to use single-trial spiking patterns to discriminate between eight unique envelope tokens (details in Materials and Methods). Classification accuracy was measured for individual cells recorded during the sinusoidal AM tokens shown in Figures 5 and for the cells recorded during the speech stimulus tokens in Figure 8. As inhibitory cells are less likely to project information to a downstream decoder, we focused on RS cells in the following analyses. A total of 16 trials were used for training the algorithm, and classification was tested based on the response to one additional trial. Note that the number of cells included in this analysis is lower ($N = 180$ RS for Sinusoidal AM; $N = 78$ RS for vocoded speech), as previous analyses required only 12 trials of each stimulus while this one requires a minimum of 17. The number of trials used in this analysis was identified by systematically determining the amount of data needed to achieve reliable templates and avoid overfitting.

Classifier results are shown in the confusion matrix in Figure 10*A*, top, for one example cell from the sinusoidal dataset (same cell as Fig. 3*A*). For this neuron, the percentage of trials that were assigned to the correct stimulus ranged from 9% to 50%. Results can also be expressed as d′ values, which incorporate the error rate for each stimulus into the metric (Fig. 10*A*, bottom). Although the PSTHs in Figure 3 show clear modulation with the AM stimuli, single trial spiking varied enough that all d′ values for this cell were under 1.2. The same classification metrics are shown for an example cell from the speech stimulus set (Fig. 10*C*, same cell as Fig. 6*A*). This cell achieved a high d′ for one stimulus, classifying 85% of trials correctly (d′ = 2.1), but several other stimuli were classified near chance levels (12.5%, d′ = 0).

Classifier results for each RS cell are plotted in Figure 10*B*. The matrix depicts d′ values on a color scale illustrating how accurately each cell classified each stimulus. Thus, each column of this matrix shows the d′ values across stimuli, like that shown for the example cell in Figure 10*A*. The same data are illustrated for the population of cells tested with vocoded speech classification (Fig. 10*D*). In each panel, a hierarchical clustering algorithm was used to sort the columns (cells) by similarity of their stimulus classification accuracies. Independently, matrix rows were clustered, which grouped stimuli with similar representations across the population of cells. A dendrogram of the stimulus clustering result is shown to the left of each matrix. In agreement with the results presented previously, sinusoidal AM rates that evoke undisputedly phase locked responses (≤ 8 Hz) group together. In other words, a cell that displays a high d′ for 4 Hz is

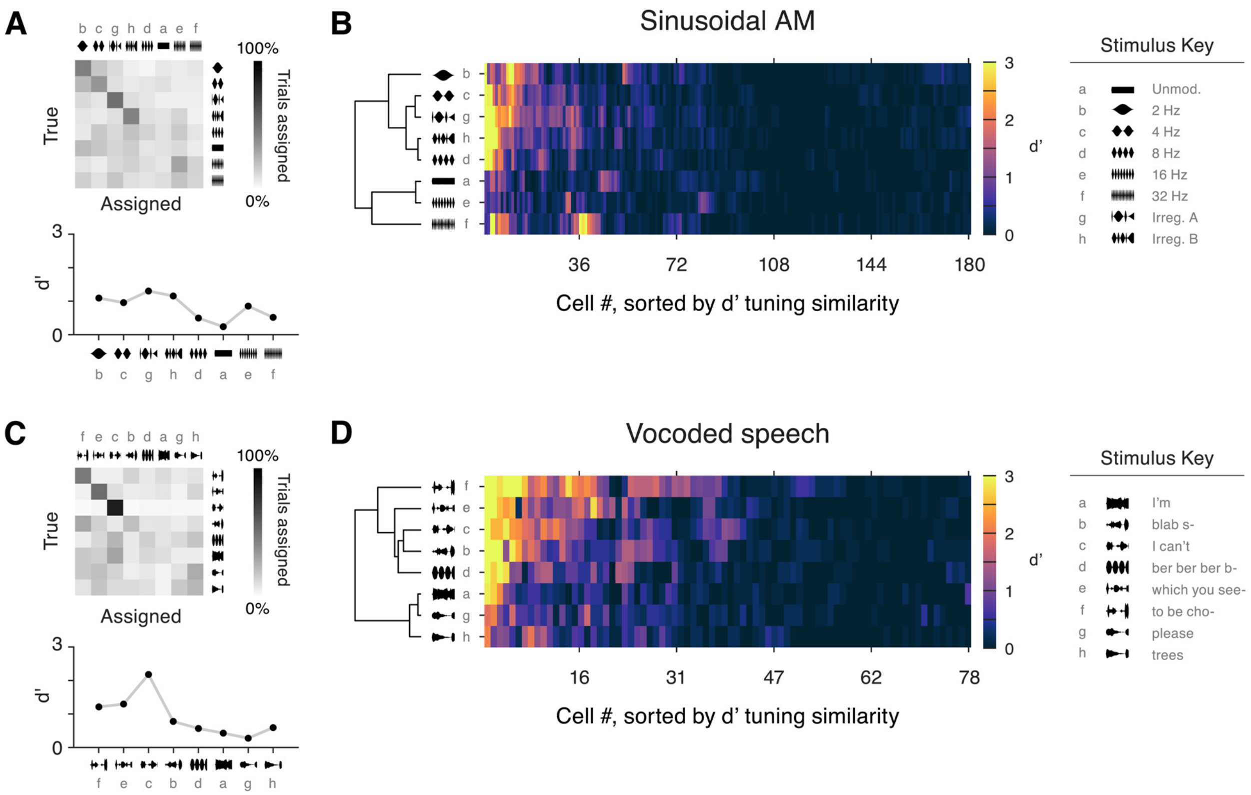

**Figure 10.** Poor envelope classification by individual neurons. ***A***, A confusion matrix displays the results of the single-trial SVM classification procedure for an example cell in the sinusoidal AM condition (same cell as Fig. 3*A*). A d′ value was calculated for each stimulus and is plotted below. Note that the order of stimulus tokens is different from prior figures; see stimulus key on far right. ***B***, d′ values across stimuli for all RS cells. Cells were sorted according to similarity of their d′ vectors using a hierarchical clustering algorithm. The same clustering approach was independently applied to sort stimulus tokens according to similarity of population representation. Overall, 9% of cell/stimulus pairs had classification levels of d′ > 1. ***C***, Results from the vocoded speech classification task are shown for an example cell (same as Fig. 6*A*). Below, d′ is shown for each stimulus. ***D***, d′ values across stimuli for all RS cells. Cells and stimuli were hierarchically clustered as above; 18% of cell/stimulus pairs had d′ > 1.

likely to decode 2 and 8 Hz with similar success. Speech stimuli also clustered into groups based on d′ values.

Overall, single-trial spiking activity from individual neurons did not allow for reliable envelope classification. The population of RS cells, on the sinusoidal AM classification task overall, had a mean d′ value of 0.32, a median d′ value of 0.11, and 13/180 units showed task classification levels of d′ > 1. Performance was similar for the vocoded speech classification task (mean d′ across cells = 0.52, median d′ across cells = 0.29, and 16/78 units with task d′ > 1). The full distribution of d′ values for each cell-stimulus pair is displayed by the ranked distributions in Figure 11*A*, black dots. Of all cell-stimulus combinations from the sinusoidal dataset, 133/1440 (9.2%) exceeded a d′ of 1, and median performance was d′ = 0.09. For the speech dataset, 18% of cell-stimulus pairs had d′ > 1, and the median d′ was 0.18 (Fig. 11*B*). These low d′ values contrast with the observation that 60–70% of cells are significantly modulated by at least one AM stimulus as measured by firing rate and/or synchronization metrics, although the classification procedure had access to all rate and temporal information (Fig. 4).

An implicit assumption in discussions of sparse coding holds that the most informative cells for a particular stimulus are those with the highest firing rates (Willmore et al., 2011; Barth and Poulet, 2012; Ince et al., 2013). Several prior studies have, in fact, identified a correlation between decoding accuracy and mean firing rate (Hoglen et al., 2018). However, the relationship is not clear-cut in all datasets, as demonstrated in Figure 11. First, all cell/stimulus pairs were ranked by d′ (Fig. 11*A*,*B*, black dots). Overlaid, we plotted a cumulative count of the average number of spikes produced during each response (Fig. 11*A*,*B*, orange line). While the number of spikes produced by a cell in response to a stimulus is significantly correlated to its decoding accuracy (Pearson, sinusoidal: $p = 2.34e^{-76}$; speech: $p = 2.18e^{-27}$), only a small fraction of the variance of d′ values was explained by the number of spikes in a response (sinusoidal AM: $r^2 = 0.21$, speech: $r^2 = 0.17$). Skewness values of the distributions of d′ values were greater than skewness of firing rates (sinusoidal AM: $\gamma_{d'} = 3.3 > \gamma_{\mathrm{Nspk}} = 2.7$; speech: $\gamma_{d'} = 2.0 > \gamma_{\mathrm{Nspk}} = 1.6$). Ultimately, 78% and 71% of spikes were produced by cells with a d′ below 1, for sinusoidal AM and vocoded speech respectively. If sound envelope were represented by a sparse coding model, it would be essential for a decoder to identify and segregate signals from the most informative cells. While there are other ways that informative cells could be identified, the present analysis suggests that it would be a nontrivial task for a downstream decoder to isolate signals from the most informative cells.

### Simple pooling supports classification of slow AM and speech

Most individual neurons could not provide reliable information about sound envelopes on single trials, so activity must be pooled across several cells to achieve envelope discrimination on par with perception. At present, there is scant evidence to inform the projection and convergence patterns within each of AC's many targets, so we assessed population coding by all RS cells. Two population decoding strategies were examined to compare envelope classification at either extreme when combining information across cells. At one extreme, inputs from cells are pooled separately and each provides independent evidence to the classifier (Fig. 12*A*,*B*, top). Independent pooling allows for maintenance of the identity (tuning preferences) of each cell, e.g., inputs arrive via separate synapses with weights that can be independently regulated. Alternatively, signals can be summed together when input to the classifier (Fig. 12*A*,*B*, bottom). Summed pooling is agnostic to the source of each spike and thus unable to weight inputs individually. However, this method requires minimal resources and requires no assumptions about the specificity of AC projections. If independent pooling were to yield better classification performance, it would suggest that envelope information is conveyed in the unique spiking patterns of individual cells (Fig. 12*A*). This would be the case if, for example, neurons produced synchronized responses with stable but heterogeneous preferred phases or response delays. Alternatively, if responses share common temporal dynamics (as indicated by previous analyses), classification based on summed population activity would be equally as good as the more sophisticated decoding strategy (Fig. 12*B*).

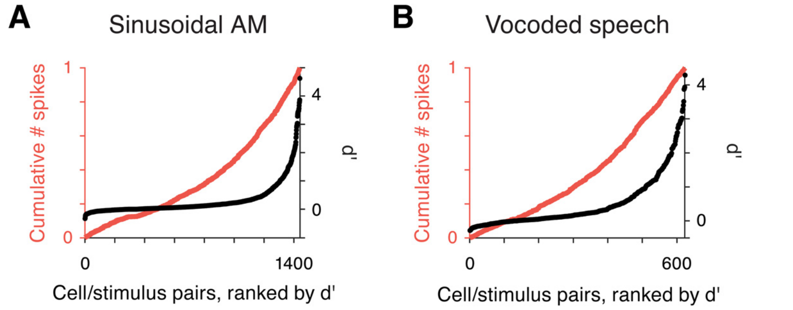

**Figure 11.** Distribution of information is sparser than distribution of spikes. ***A***, The ranked distribution of d′ values is plotted for all cell/stimulus pairs included in the sinusoidal AM stimulus classification analyses (black dots, right vertical axis). Overlaid is a cumulative count of the number of spikes in the responses of the corresponding cell/stimulus pairs (orange line, left vertical axis). The skewness of the distribution of d′ values is higher than the skewness of the number of spikes ($\gamma_{d'} = 3.3 > \gamma_{\mathrm{Nspk}} = 2.7$). ***B***, Ranked d′ values are plotted for all cell/stimulus pairs for vocoded speech classification (black dots). The corresponding number of spikes in the response plotted as a cumulative distribution (orange line). As above, the distribution of spikes is less skewed than that of d′ values ($\gamma_{d'} = 2.0 > \gamma_{\mathrm{Nspk}} = 1.6$).

We compared the results of two population classifiers designed to approximate the integration strategies described above, to gain a sense of how simple the decoding rule could be. Specifically, does the collective activity of the population suffice for detecting prominent edges in the envelope signal? For the sinusoidal AM dataset, the average classification accuracy was d′ = 3.98 when information from all RS cells in the population was included independently (Fig. 12*C*, left). When spiking activity was summed across cells before being passed to the classifier, average performance for the task dropped to d′ = 2.15, suggesting that independent integration is advantageous. However, when classification results were examined for each stimulus separately, it was apparent that some AM rates benefited from maintaining independent inputs more than others (Fig. 12*C*, right). Stimuli containing slower AM rates (2 Hz, 4 Hz, and the irregular segment containing a 4-Hz period) showed roughly equivalent classification performance when activity was summed across cells.

To estimate the best performance that could be achieved from this population, we gradually pooled cells from best to worst individual d′ (Fig. 12*D*). Classification based on independent integration required only a few of the best neurons to reach d′ = 4. When decoding was measured from the summed activity pattern, performance was best with a selective ensemble of the best cells, and d′ converged to roughly the level of the best individual unit when all cells in the population were included.

Which population decoding model is better for decoding the envelopes of human speech? Would a decoder need to maintain the identity of individual cells, or is the temporal pattern of activity in the global population sufficient? On average for the task, the two pooling methods result in similar d′ values (Fig. 12*E*, left). Unlike most sinusoidal AM stimuli, each individual speech stimulus demonstrated similar decoding accuracy from summed and independent pooling methods (Fig. 12*E*, right). Additionally, the two decoding methods show consistent performance across ensembles of varying sizes (Fig. 12*F*). As before, performance of the entire RS population converges to approximately the level of the best individual cell.

Overall, decoding syllabic structure does not require independently tuned inputs. Although tuning and responses across the AC population are heterogeneous, our results suggest that the onset edges of sound amplitude coherently shape the temporal response dynamics in a manner that creates an easily decoded population level signal.

## Narrow spiking cells

Our analyses have focused primarily on stimulus encoding and decoding by RS cells, which are thought to represent the principal neurons that project to downstream populations. However, we also recorded from a smaller number of narrow spiking (NS) cells, thought to reflect inhibitory interneurons (Wilson et al., 1994; Barthó et al., 2004; Mesik et al., 2015; F. Liang et al., 2019). The population of NS cells displayed higher spontaneous and evoked (Fig. 4*A*) firing rates as compared with RS cells, suggesting that our recordings primarily sampled fast spiking interneurons (Li et al., 2015).

To compare the envelope-evoked responses of RS and NS cells, we plotted the normalized population activity surrounding peak derivative events in the envelope signal for each AM rate (Fig. 13*A*). The temporal pattern of spiking differed for RS and NS cells. Most notably, NS cell activity dipped around 40–50 ms after a peakDrv event, coinciding with the time that RS activity peaked. This discharge pattern was particularly apparent for slower AM rates, suggesting that a brief period of enhanced excitatory transmission follows the onset edges in the amplitude envelope signal. Speech stimuli followed this pattern as well: the peak in RS firing occurs when NS activity is at a minimum (Fig. 13*B*). However, our ability to draw quantitative conclusions about NS cells' vocoded speech encoding is limited by the relatively few NS units recorded in this condition ($N = 11$).

Since a small percentage of inhibitory neurons do project over long distances (Melzer and Monyer, 2020; Urrutia-Piñones et al., 2022), we also evaluated the envelope decoding properties of NS cells for the sinusoidal AM stimulus tokens (Fig. 13*C*). The distribution of single-trial decoding accuracies was similar to RS cells: 11.1% of NS cell/stimulus pairs achieved d′ > 1 (as compared with 9.2% of RS cell/stimulus pairs). Like RS cells, only a small amount of the variance in d′ values was explained by the number of spikes in a response (sinusoidal AM: $r^2 = 0.04$). In contrast to RS cells, however, NS cells demonstrated more accurate decoding of faster AM rates (32 Hz d′ > 1: 25.9% of NS

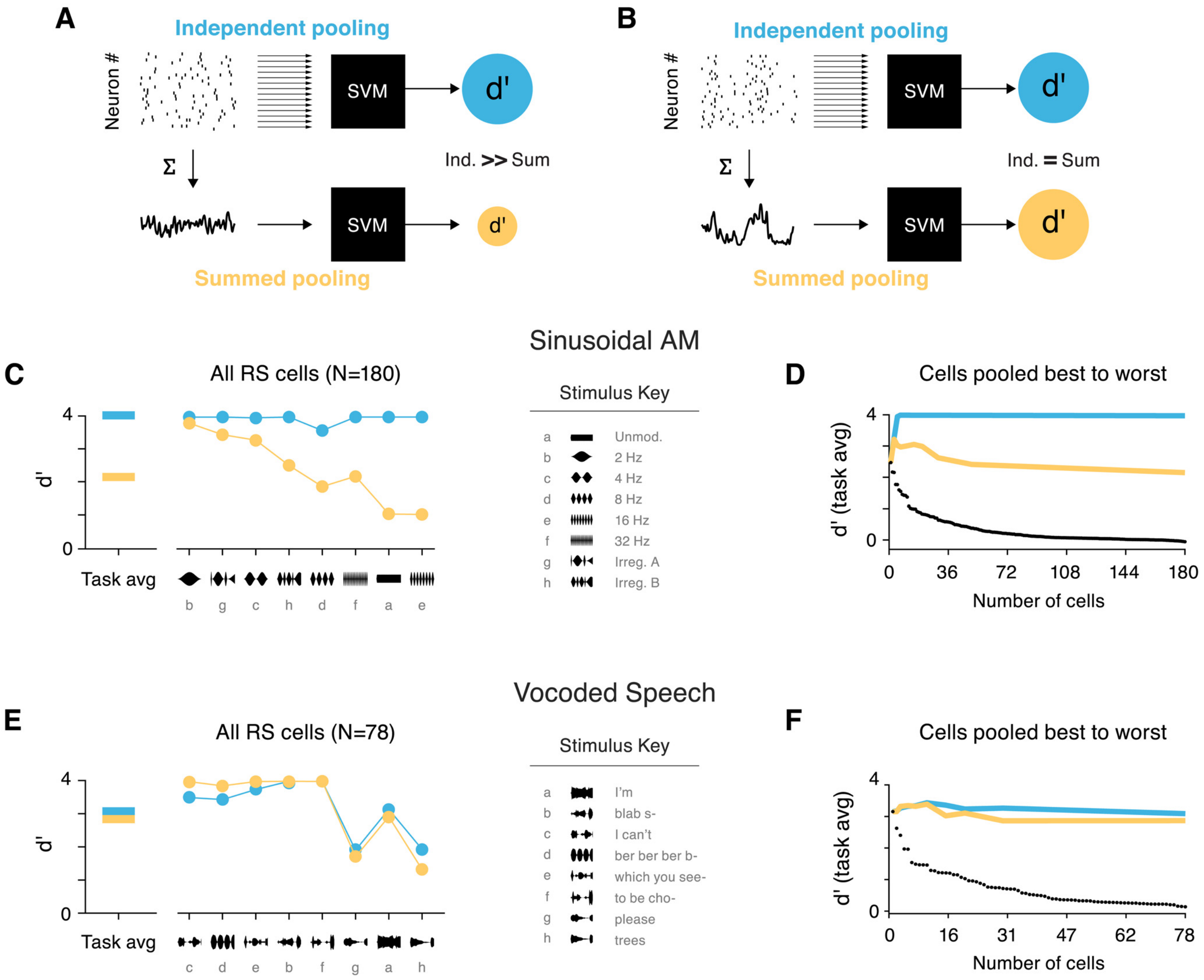

**Figure 12.** Syllable-rate envelopes can be decoded from the sum of population activity. Comparison of two population decoding methods based on single-trial data. ***A***, Schematic shown for a case where independent pooling (blue) yields better decoding accuracy than summed pooling (yellow). $d'_{\text{Ind}}$ (top, blue) results from classification based on the single trial spiking data from each RS cell in the population. $d'_{\text{Sum}}$ (bottom, yellow) collapses spiking activity into one vector, and SVM classification is based on the temporal profile of activity in the population. In this example, $d'_{\text{Ind}} > d'_{\text{Sum}}$ because the information carried by individual cells does not manifest as a common temporal pattern of spiking. ***B***, Decoding schematic shown for a case where $d'_{\text{Sum}} = d'_{\text{Ind}}$. Here, temporal structure related to the envelope is preserved when activity is summed across the population. ***C***, Classifier results, for the Sinusoidal AM task (left) and broken down by stimulus (right), from independent pooling (blue) and summed pooling (yellow). Stimuli are sorted in ascending order of $d'_{\text{Ind}}–d'_{\text{Sum}}$. The three stimuli with prominent 2- or 4-Hz periods show $d'_{\text{Sum}}$ performs nearly as well as $d'_{\text{Ind}}$, while faster AM rates benefit from independent integration of cells. ***D***, Average classification in the sinusoidal AM task for each pooling method, with increasing ensemble cells. Cells are gradually added to the population from best to worst d′ (black dots). Overall, independent pooling is needed to sufficiently discriminate between all stimuli in this set. Performance from pooling all cells is roughly equivalent to performance when including only a few of the best neurons. ***E***, Results for the vocoded speech stimulus set: task average (left) and broken down by stimulus (right), presented as in panel ***C***. Performance is equivalent for both decoding methods, meaning that independent integration of cells' activity is not required to distinguish between speech envelopes. ***F***, Classification with increasing pool size, as presented in panel ***D***. Average performance from the two decoding methods is similar across all ensemble sizes.

cells vs 10.6% of RS cells). This observation is consistent with the results of Figure 4C, where a greater fraction of NS cells synchronized to higher modulation rates according to vector strength metrics.

## Discussion

Animal communication sounds have rich temporal structure and are often produced in extended sequences of "syllables," or packets of information, occurring at rates of ~2–8 Hz. Human speech research has built a strong case that AC plays a crucial role in comprehension, in part, by segmenting the acoustic stream into syllables (Giraud and Poeppel, 2012; Peelle and Davis, 2012; Haegens and Zion Golumbic, 2018). The cellular mechanisms underlying this parsing operation are unclear because single neuron responses are largely studied during periodic AM stimuli and because traditional analyses do not examine the absolute temporal relationship between spikes and the stimulus. In the present study, we measured responses of a broad population of AC neurons and asked how the discharge patterns of individual neurons contribute to envelope representations of complex, natural sounds like speech. While AM rate and synchronization tuning of neurons was heterogeneous, the timing of responses was stereotyped across cells. We found that preferred phases collected near the onsets of sinusoidal AM periods, corresponding to transient increases in the global

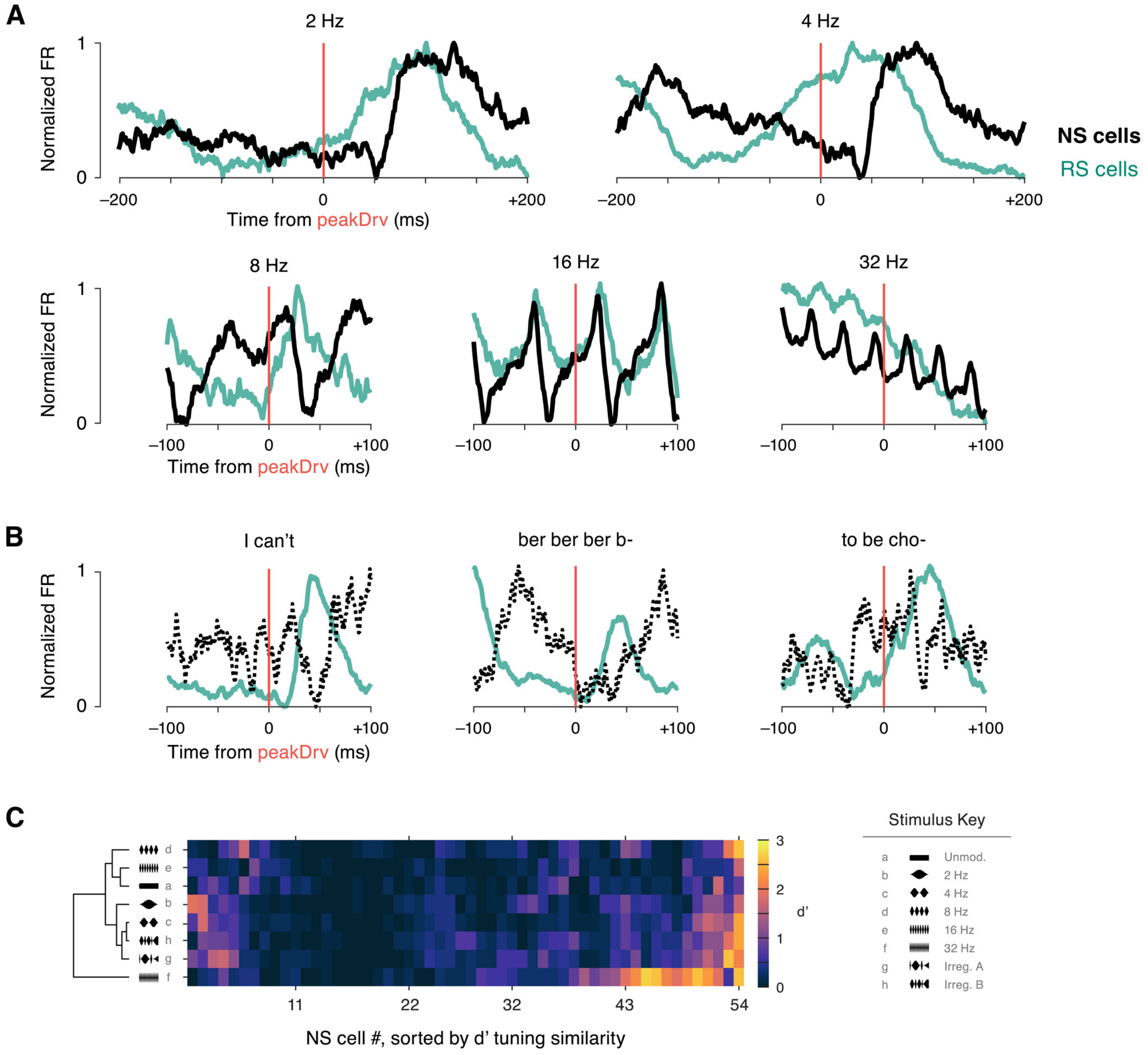

**Figure 13.** Narrow spiking (NS) cells display a temporally-distinct response pattern and similar single-unit decoding accuracy. ***A***, The evoked response surrounding peakDrv landmarks in each periodic sinusoidal AM stimulus is shown for the RS population (green, $N = 181$) and the NS population (black, $N = 54$). Analysis is identical to Figure 9, but population mean firing rates are normalized to facilitate comparison between cell types. Note that NS cell firing dips around 40- to 50-ms latency, at the same time RS cell firing peaks. ***B***, Evoked responses for RS cells ($N = 100$) and NS cells ($N = 11$) are shown for three example vocoded speech stimuli. Only preliminary observations can be made about NS population activity (dotted line) because of the limited number of cells recorded with at least 12 trials for each vocoded speech stimulus. Still, the same pattern is visible, with NS firing decreasing and RS firing increasing at short latency following peakDrv events. ***C***, Single trial decoding accuracy is shown for each NS cell. d′ values are illustrated by color scale (right) for each NS cell (columns) and each sinusoidal AM stimulus (rows). As in Figure 10, rows and columns are independently sorted by hierarchical clustering. Overall, 11.1% of NS cell/stimulus pairs had classification levels of $d' > 1$, compared with 9.2% for RS cells. NS cells were more successful at decoding the 32-Hz stimulus (25.9% of NS cells with $d' > 1$, compared with 10.6% of RS cells with $d' > 1$).

population signal marking amplitude edges. Only modulation rates in the range of natural syllables exhibited this pattern, although phase-locked responses continued at higher modulation rates (Figs. 5, 9). The stimuli that displayed an overrepresentation of onset edges in encoding analyses also demonstrated an advantage for decoding: while classification of higher frequency modulations required input from individual cells, vocoded speech and syllable-like stimuli were discriminated equally well using the summed population activity from single trials (Fig. 12). These results suggest that a phasic, redundant code in the AC population provides a mechanism for segmenting acoustic streams like human speech.

### Distributed, coherent spiking encodes amplitude edges

Prior work on AM encoding established diversity across cells as a principal feature of sinusoidal AM rate tuning. This pattern has been observed regardless of whether the carrier was broadband noise or tones optimized for each cell individually, and regardless of whether responses were measured by firing rate, temporal metrics, or a classifier (Schreiner and Urbas, 1988; Eggermont, 1998; L. Liang et al., 2002; Joris et al., 2004; Malone et al., 2007; Zhou and Wang, 2010; Yin et al., 2011; Hoglen et al., 2018). Heterogeneity has also been reported when analyzing the shapes of modulation period histograms in the inferior colliculus (Rees and Møller, 1983; Krishna and Semple, 2000) and in AC (Malone et al., 2007).

Temporal responses to sinusoidal AM are measured by the mean phase of spikes. However, this metric is typically reported only for example neurons or is inspected on a within-cell basis (Eggermont, 1998). In this study, we directly quantified the temporal relationship between spiking across the AC population and both simple and complex envelopes. Despite a strong precedent of heterogeneity, we found that mean phases were heavily biased to occur in the first 90° for AM rates in the fluctuation range (Fig. 5). Our results are consistent with a recent study of single-unit sinusoidal AM responses in squirrel monkeys, which reported shared phase preferences across neurons in AC and high accuracy of a population-based decoding model (Downer et al., 2021). Their results suggested that population coding of speech envelopes would be robust to indiscriminate pooling of cortical responses. Our results obtained with vocoded speech confirm that prediction (Fig. 12). In addition, RS neurons (putative projection neurons) and NS neurons (putative inhibitory interneurons) were cleanly distinguished in our dataset. Both displayed phasic AM responses, and RS cells exhibited a bias toward increased spiking following amplitude edges, at a time when inhibition is briefly reduced.

If individual RS neurons contributed differential information to the envelope representation, summing activity across cells would detract from classification performance. In contrast, we found a classifier could extract sufficient information to discriminate envelope stimuli using a single-trial population activity vector, with no benefit of tracking inputs from individual cells. Natural speech contains substantially more complexity than one-channel vocoded speech. Instead of a single envelope signal, envelope cues involve correlated modulations in amplitude across frequency bands. Thus, envelopes of complex, natural sounds may involve distributed codes within a few subpopulations of neurons simultaneously, in addition to the sparse coding that is likely needed to decode spectral information.

## Relationship between NS and RS properties

Thus far, we have focused on regular spiking (RS) cells which are thought to reflect principal neurons that serve as the primary output of AC to downstream regions. However, we also characterized a population of narrow spiking (NS) cells that are thought to reflect inhibitory interneurons (Wilson et al., 1994; Barthó et al., 2004; Mesik et al., 2015; F. Liang et al., 2019). Functionally, inhibitory interneurons are largely involved in local networks that sculpt RS cell coding properties, and they are also implicated in a broad range of AC coding properties (Tsunada et al., 2012; Natan et al., 2015, 2017; Seay et al., 2020; Pérez-González et al., 2021), as well as long-term auditory plasticity (Jeanne et al., 2013; Sarro et al., 2015; Resnik and Polley, 2017; Vickers et al., 2018; Mowery et al., 2019). That said, we acknowledge that there are many types of cortical interneurons (Kepecs and Fishell, 2014) including a small percentage of GABAergic neurons with long-range projections (Melzer and Monyer, 2020; Urrutia-Piñones et al., 2022), and spike kinetics cannot conclusively determine cell types (Moore and Wehr, 2013).

In layers 2–4, the primary location of our recordings (Fig. 2; Materials and Methods), principal neurons are known to integrate at least two, functionally distinct, inhibitory inputs (Beierlein et al., 2003; Tan et al., 2008; Cruikshank et al., 2010). In gerbil auditory cortex, fast-spiking interneurons provide feed-forward inhibition, and exhibit large, reliable inhibitory potentials that display short-term depression, while low-threshold spiking interneurons produce smaller inhibitory potentials that do not display as much synaptic depression (Takesian et al., 2010, 2013).

The population of NS cells reported in the present study displayed high spontaneous and evoked firing rates (Figs. 1*F*, 4*A*), and phasic responses that continued at higher AM rates than RS cells (Fig. 4*C*), consistent with previous descriptions of fast spiking inhibitory interneurons in AC (Atencio and Schreiner, 2008; Levy and Reyes, 2012; Moore and Wehr, 2013; Li et al., 2015; Bottjer et al., 2019; Gao and Wang, 2019; Liu and Wang, 2022). Furthermore, we found that the relative response latencies of NS cells, as referenced to the peaks in the derivative of the envelope signal (peakDrv), were longer than those of RS cells at slow modulation rates (Fig. 13*A*). In fact, NS cell firing briefly decreased at the time of the RS cell peak response, suggesting that NS cells constrain the RS response window. This is consistent with experimental evidence that parvalbumin-positive (PV+) interneurons shape RS cell discharge pattern. For example, pharmacological manipulations of GABAergic transmission in gerbil AC (Kurt et al., 2006) or optogenetic manipulations of PV+ cells in mouse AC brain slices (Krause et al., 2019) result in selective alterations to RS cell discharge patterns.

In the present study, latencies of NS cell peak firing following onsets of amplitude edges often occurred later than RS cell peaks. This observation contrasts with reports from PV+ fast-spiking neurons in mouse AC, which display shorter latencies than non-PV cells in response to medial geniculate stimulation or to pure tones (Rose and Metherate, 2005; Moore and Wehr, 2013). A possible reconciliation of these observations is that continuous sound streams may lead to different spectrotemporal adaptation and circuit dynamics as compared with isolated stimuli with rapid rise times (e.g., tone pulses). Future studies should continue to probe how envelope coding is sculpted by the relative timing of excitation and inhibition and adaptation to ongoing sound stimuli.

## Limitations of the dataset

While our recordings spanned the tonotopic axis of AC, we did not directly assess spectral tuning. However, there is reason to believe that envelope representation would be robust to changes in spectral parameters. Global, population-level signals in human AC and single units in animals have been shown to encode amplitude envelopes independently from spectral features (Malone et al., 2007; Oganian and Chang, 2019). Previous experiments using pure tone carriers also predicted that amplitude edge detection could be accomplished by a collective signal from the AC population (Zhou and Wang, 2010; Downer et al., 2021). Further, behavioral experiments using chimerized stimuli, which dissociate envelope and carrier components, provide perceptual evidence of independent processing (Smith et al., 2002).

Additionally, the upper bound of modulation rates for which a redundant code exists is not clear. In the present study, the range of phase locking extended higher than the range for which coherent phase preferences were observed. Studies in different species reveal different synchronization limits of single units in AC, with nonhuman primates showing phase-locking retained at faster modulation rates than rodent species (Hoglen et al., 2018). The results in Downer et al. (2021), obtained from squirrel monkeys, demonstrate that coherent phase preferences stretch to higher modulation rates in species with higher synchronization limits. Still, future studies will have to investigate the relationship between synchronization and phase preferences, how these properties influence population coding, and whether these properties extend to single neurons in human auditory cortex.

### Limitations of a redundant code

The present evidence suggests that a distributed, redundant code might be available and useful for the task of parsing a continuous signal like speech. Faster modulation rates (16 and 32 Hz in the present dataset) did not exhibit the same representation, suggesting that this redundant code would not contribute to perception of acoustic elements such as timbre and pitch although these properties also rely on envelope cues. Similarly, discriminating between similar modulation shapes may also require higher resolution neural representations. For example, the classifier used in this study struggled to discriminate the vocoded speech segments corresponding to the words "please" and "trees" (Figs. 10, 12). A similar observation was made in a study assessing envelope responses at a different scale, using ECoG recordings: responses were qualitatively identical for linear-shaped and nonlinearly-shaped amplitude ramps (Oganian and Chang, 2019).

While we have emphasized the similarity of cells' responses, substantial richness and diversity did exist in the population (Figs. 4, 5, 8). This heterogeneity reflects the tuning specialization that is characteristic of cortex and is valuable for coding many other key sound dimensions. Tuning specificity is integral in sparse coding models, which have been shown to be effective for discriminating natural sound stimuli that vary greatly in spectral characteristics (Ince et al., 2013; Schneider and Woolley, 2013). Sparse and redundant codes are not mutually exclusive, though; they can exist in parallel. This phenomenon was demonstrated in piriform cortex: the identity of active neurons conveyed odor identity, while a distributed temporal code represented odor intensity (Bolding and Franks, 2017). Future experiments should be designed to look for multiplexed acoustic information in AC.

There is reason to believe that multiplexing could occur in auditory cortex. For example, transient responses are observed at sound onset (from silence) in marmoset neurons with a variety of tuning preferences, but sustained responses occur only in a subset of cells precisely-tuned to the current acoustic parameters (Wang et al., 2005). Note that the prevalence of sustained responses likely depends on species: 43% of rhesus neurons display a sustained response to long pure tone stimuli (Malone et al., 2007). Regardless, the present results suggest that relatively promiscuous onset responses extend beyond sound onsets from silence and also occur for amplitude edges during continuously modulating sounds. Thus, population sparseness and decoding conditions may vary over time, influenced by stimulus dynamics, allowing for multiple codes to coexist.

### Relationship to speech processing

In humans, neural recordings during both passive and active speech listening demonstrate precise tracking of speech envelopes by cortical activity, known as rhythmic tracking or entrainment (Giraud and Poeppel, 2012; Peelle and Davis, 2012; Haegens and Zion Golumbic, 2018). Intelligibility of speech depends on the vitality of phasic responses in AC; for example, microstimulation in Heschl's gyrus applied at acoustic edges disrupts comprehension (Forseth et al., 2020). However, rhythmic tracking activity is also evoked in the absence of intelligibility, suggesting a purely acoustic, or bottom-up, component of speech processing (Peelle et al., 2013; Meyer et al., 2017). Our results align with studies of speech and AM encoding in humans, which find that population activity in AC is sensitive to edges in the amplitude envelope of speech stimuli (Oganian and Chang, 2019; Forseth et al., 2020). Importantly, these acoustic edges correlate with the onsets of vowel nuclei, which perceptually define syllabic rate (Oganian and Chang, 2019). The evidence for a direct relationship between the acoustic signal and the linguistic structure of speech emphasizes the possibility that some mechanisms underlying speech processing may be studied using animal models.

The scales of neural activity in human speech studies and animal electrophysiology are vastly different, as are the complexity and salience of species' vocalizations. Is it fair to compare these data? One previous study provides insight into how cellular electrophysiology could be linked to speech parsing models in humans. Szymanski et al. (2011) presented spectrotemporally complex stimuli, including rock music, to anesthetized rats while measuring current source density (CSD) across layers of AC. Discrete events were identified in the CSD signal that reflected "high-amplitude neuronal discharge," thought to be driven by thalamocortical input. These events occurred at a rate of 2–4 Hz and, although their temporal relationship to the stimulus was not quantified, qualitatively appear correlated with the onsets of prominent amplitude edges in the stimulus. Crucially, the occurrence of high-amplitude discharge events corresponded to phase resets of low frequency local field potentials. In human speech literature, oscillatory phase resetting is a prominent mechanism implicated in speech parsing (Peelle and Davis, 2012; Haegens and Zion Golumbic, 2018). Mechanisms of phase resetting are ripe for investigation at the level of single cells and circuits in awake animals (Guo et al., 2017). Overall, by combining several levels of inquiry, we can elucidate how the information-processing operations performed by cells and circuits underlie the global signals measured in humans, helping to crack the neural mechanisms of speech processing.

## References

Atencio CA, Schreiner CE (2008) Spectrotemporal processing differences between auditory cortical fast-spiking and regular-spiking neurons. J Neurosci 28:3897–3910.

Barth AL, Poulet JFA (2012) Experimental evidence for sparse firing in the neocortex. Trends Neurosci 35:345–355.

Barthó P, Hirase H, Monconduit L, Zugaro M, Harris KD, Buzsáki G (2004) Characterization of neocortical principal cells and interneurons by network interactions and extracellular features. J Neurophysiol 92:600–608.

Beierlein M, Gibson JR, Connors BW (2003) Two dynamically distinct inhibitory networks in layer 4 of the neocortex. J Neurophysiol 90:2987–3000.

Bolding KA, Franks KM (2017) Complementary codes for odor identity and intensity in olfactory cortex. Elife 6:e22630.

Bottjer SW, Ronald AA, Kaye T (2019) Response properties of single neurons in higher level auditory cortex of adult songbirds. J Neurophysiol 121:218–237.

Cruikshank SJ, Urabe H, Nurmikko AV, Connors BW (2010) Pathway-specific feedforward circuits between thalamus and neocortex revealed by selective optical stimulation of axons. Neuron 65:230–245.

David SV, Mesgarani N, Fritz JB, Shamma SA (2009) Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. J Neurosci 29:3374–3386.

Ding N, Patel AD, Chen L, Butler H, Luo C, Poeppel D (2017) Temporal modulations in speech and music. Neurosci Biobehav Rev 81:181–187.

Downer JD, Bigelow J, Runfeldt MJ, Malone BJ (2021) Temporally precise population coding of dynamic sounds by auditory cortex. J Neurophysiol 126:148–169.

Drullman R, Festen JM, Plomp R (1994a) Effect of reducing slow temporal modulations on speech reception. J Acoust Soc Am 95:2670–2680.

Drullman R, Festen JM, Plomp R (1994b) Effect of temporal envelope smearing on speech reception. J Acoust Soc Am 95:1053–1064.

Eggermont JJ (1998) Representation of spectral and temporal sound features in three cortical fields of the cat. Similarities outweigh differences. J Neurophysiol 80:2743–2764.

Forseth KJ, Hickok G, Rollo PS, Tandon N (2020) Language prediction mechanisms in human auditory cortex. Nat Commun 11:5240.

Gao L, Wang X (2019) Subthreshold activity underlying the diversity and selectivity of the primary auditory cortex studied by intracellular recordings in awake marmosets. Cereb Cortex 29:994–1005.

Ghitza O (2012) On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. Front Psychol 3:238.

Giraud AL, Poeppel D (2012) Cortical oscillations and speech processing: emerging computational principles and operations. Nat Neurosci 15:511–517.

Green DM, Swets JA (1966) Signal detection theory and psychophysics. New York: Wiley.

Guo W, Clause AR, Barth-Maron A, Polley DB (2017) A corticothalamic circuit for dynamic switching between feature detection and discrimination. Neuron 95:180–194.e5.

Haegens S, Zion Golumbic E (2018) Rhythmic facilitation of sensory processing: a critical review. Neurosci Biobehav Rev 86:150–165.

Hoglen NEG, Larimer P, Phillips EAK, Malone BJ, Hasenstaub AR (2018) Amplitude modulation coding in awake mice and squirrel monkeys. J Neurophysiol 119:1753–1766.

Hromádka T, DeWeese MR, Zador AM (2008) Sparse representation of sounds in the unanesthetized auditory cortex. PLoS Biol 6:e16.

Ince RAA, Panzeri S, Kayser C (2013) Neural codes formed by small and temporally precise populations in auditory cortex. J Neurosci 33:18277–18287.

Ineichen BV, Weinmann O, Good N, Plattner PS, Wicki C, Rushing EJ, Linnebank M, Schwab ME (2017) Sudan black: a fast, easy and non-toxic method to assess myelin repair in demyelinating diseases. Neuropathol Appl Neurobiol 43:242–251.

Jeanne JM, Sharpee TO, Gentner TQ (2013) Associative learning enhances population coding by inverting interneuronal correlation patterns. Neuron 78:352–363.

Joris PX, Schreiner CE, Rees A (2004) Neural processing of amplitude-modulated sounds. Physiol Rev 84:541–577.

Kepecs A, Fishell G (2014) Interneuron cell types are fit to function. Nature 505:318–326.

Krause BM, Murphy CA, Uhlrich DJ, Banks MI (2019) PV+ cells enhance temporal population codes but not stimulus-related timing in auditory cortex. Cereb Cortex 29:627–647.

Krishna BS, Semple MN (2000) Auditory temporal processing: responses to sinusoidally amplitude-modulated tones in the inferior colliculus. J Neurophysiol 84:255–273.

Kurt S, Crook JM, Ohl FW, Scheich H, Schulze H (2006) Differential effects of iontophoretic in vivo application of the GABA(A)-antagonists bicuculline and gabazine in sensory cortex. Hear Res 212:224–235.

Levy RB, Reyes AD (2012) Spatial profile of excitatory and inhibitory synaptic connectivity in mouse primary auditory cortex. J Neurosci 32:5609–5619.

Li LY, Xiong XR, Ibrahim LA, Yuan W, Tao HW, Zhang LI (2015) Differential receptive field properties of parvalbumin and somatostatin inhibitory neurons in mouse auditory cortex. Cereb Cortex 25:1782–1791.

Liang F, Li H, Chou XL, Zhou M, Zhang NK, Xiao Z, Zhang KK, Tao HW, Zhang LI (2019) Sparse representation in awake auditory cortex: cell-type dependence, synaptic mechanisms, developmental emergence, and modulation. Cereb Cortex 29:3796–3812.

Liang L, Lu T, Wang X (2002) Neural representations of sinusoidal amplitude and frequency modulations in the primary auditory cortex of awake primates. J Neurophysiol 87:2237–2261.

Liu XP, Wang X (2022) Distinct neuronal types contribute to hybrid temporal encoding strategies in primate auditory cortex. PLoS Biol 20: e3001642.

Malone BJ, Scott BH, Semple MN (2007) Dynamic amplitude coding in the auditory cortex of awake rhesus macaques. J Neurophysiol 98:1451–1474.

Malone BJ, Beitel RE, Vollmer M, Heiser MA, Schreiner CE (2013) Spectral context affects temporal processing in awake auditory cortex. J Neurosci 33:9431–9450.

Malone BJ, Beitel RE, Vollmer M, Heiser MA, Schreiner CE (2015) Modulation-frequency-specific adaptation in awake auditory cortex. J Neurosci 35:5904–5916.

Melzer S, Monyer H (2020) Diversity and function of corticopetal and corticofugal GABAergic projection neurons. Nat Rev Neurosci 21:499–515.

Mesik L, Ma W, Li L, Ibrahim LA, Huang ZJ, Zhang LI, Tao HW (2015) Functional response properties of VIP-expressing inhibitory neurons in mouse visual and auditory cortex. Front Neural Circuits 9:22.

Meyer L, Henry MJ, Gaston P, Schmuck N, Friederici AD (2017) Linguistic bias modulates interpretation of speech via neural delta-band oscillations. Cereb Cortex 27:4293–4302.

Moore AK, Wehr M (2013) Parvalbumin-expressing inhibitory interneurons in auditory cortex are well-tuned for frequency. J Neurosci 33:13713–13723.

Mowery TM, Caras ML, Hassan SI, Wang DJ, Dimidschstein J, Fishell G, Sanes DH (2019) Preserving inhibition during developmental hearing loss rescues auditory learning and perception. J Neurosci 39: 8347–8361.

Natan RG, Briguglio JJ, Mwilambwe-Tshilobo L, Jones SI, Aizenberg M, Goldberg EM, Geffen MN (2015) Complementary control of sensory adaptation by two types of cortical interneurons. Elife 4:e09868.

Natan RG, Rao W, Geffen MN (2017) Cortical interneurons differentially shape frequency tuning following adaptation. Cell Rep 21:878–890.

Oganian Y, Chang EF (2019) A speech envelope landmark for syllable encoding in human superior temporal gyrus. Sci Adv 5:eaay6279.

Peelle JE, Davis MH (2012) Neural oscillations carry speech rhythm through to comprehension. Front Psychol 3:320.

Peelle JE, Gross J, Davis MH (2013) Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb Cortex 23:1378–1387.

Pérez-González D, Parras GG, Morado-Díaz CJ, Aedo-Sánchez C, Carbajal GV, Malmierca MS (2021) Deviance detection in physiologically identified cell types in the rat auditory cortex. Hear Res 399:107997.

Radtke-Schuller S, Schuller G, Angenstein F, Grosser OS, Goldschmidt J, Budinger E (2016) Brain atlas of the Mongolian gerbil (*Meriones unguiculatus*) in CT/MRI-aided stereotaxic coordinates. Brain Struct Funct 221:1–272.

Rees A, Møller AR (1983) Responses of neurons in the inferior colliculus of the rat to AM and FM tones. Hear Res 10:301–330.

Resnik J, Polley DB (2017) Fast-spiking GABA circuit dynamics in the auditory cortex predict recovery of sensory processing following peripheral nerve damage. Elife 6:e21452.

Rose HJ, Metherate R (2005) Auditory thalamocortical transmission is reliable and temporally precise. J Neurophysiol 94:2019–2030.

Sadagopan S, Wang X (2009) Nonlinear spectrotemporal interactions underlying selectivity for complex sounds in auditory cortex. J Neurosci 29:11192–11202.

Sarro EC, von Trapp G, Mowery TM, Kotak VC, Sanes DH (2015) Cortical synaptic inhibition declines during auditory learning. J Neurosci 35:6318–6325.

Schneider DM, Woolley SMN (2013) Sparse and background-invariant coding of vocalizations in auditory scenes. Neuron 79:141–152.

Schreiner CE, Urbas JV (1988) Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between cortical fields. Hear Res 32:49–63.

Seay MJ, Natan RG, Geffen MN, Buonomano DV (2020) Differential short-term plasticity of PV and SST neurons accounts for adaptation and facilitation of cortical neurons to auditory tones. J Neurosci 40: 9224–9235.

Singh NC, Theunissen FE (2003) Modulation spectra of natural sounds and ethological theories of auditory processing. J Acoust Soc Am 114:3394–3411.

Smith ZM, Delgutte B, Oxenham AJ (2002) Chimaeric sounds reveal dichotomies in auditory perception. Nature 416:87–90.

Szymanski FD, Rabinowitz NC, Magri C, Panzeri S, Schnupp JWH (2011) The laminar and temporal structure of stimulus information in the phase of field potentials of auditory cortex. J Neurosci 31: 15787–15801.

Takesian AE, Kotak VC, Sanes DH (2010) Presynaptic GABA(B) receptors regulate experience-dependent development of inhibitory short-term plasticity. J Neurosci 30:2716–2727.

Takesian AE, Kotak VC, Sharma N, Sanes DH (2013) Hearing loss differentially affects thalamic drive to two cortical interneuron subtypes. J Neurophysiol 110:999–1008.

Tan Z, Hu H, Huang ZJ, Agmon A (2008) Robust but delayed thalamocortical activation of dendritic-targeting inhibitory interneurons. Proc Natl Acad Sci U S A 105:2187–2192.

Tsunada J, Lee JH, Cohen YE (2012) Differential representation of auditory categories between cell classes in primate auditory cortex. J Physiol 590:3129–3139.

Urrutia-Piñones J, Morales-Moraga C, Sanguinetti-González N, Escobar AP, Chiu CQ (2022) Long-range gabaergic projections of cortical origin in brain function. Front Syst Neurosci 16:841869.

Vickers ED, Clark C, Osypenko D, Fratzl A, Kochubey O, Bettler B, Schneggenburger R (2018) Parvalbumin-interneuron output synapses show spike-timing-dependent plasticity that contributes to auditory map remodeling. Neuron 99:720–735.e6.

Wang X, Lu T, Snider RK, Liang L (2005) Sustained firing in auditory cortex evoked by preferred stimuli. Nature 435:341–346.

Willmore BDB, Mazer JA, Gallant JL (2011) Sparse coding in striate and extrastriate visual cortex. J Neurophysiol 105:2907–2919.

Wilson FA, O'Scalaidhe SP, Goldman-Rakic PS (1994) Functional synergism between putative gamma-aminobutyrate-containing neurons and pyramidal neurons in prefrontal cortex. Proc Natl Acad Sci U S A 91:4009–4013.

Yin P, Johnson JS, O'Connor KN, Sutter ML (2011) Coding of amplitude modulation in primary auditory cortex. J Neurophysiol 105: 582–600.

Zhou Y, Wang X (2010) Cortical processing of dynamic sound envelope transitions. J Neurosci 30:16741–16754.